Visual Analysis in Traffic & Re-identification

Møgelmose, Andreas

DOI (link to publication from Publisher):
10.5278/vbn.phd.engsci.00026

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

VISUAL ANALYSIS IN TRAFFIC & RE-IDENTIFICATION

BY
ANDREAS MØGELMOSE

DISSERTATION SUBMITTED 2015

AALBORG UNIVERSITY
DENMARK

Visual Analysis in Traffic & Re-identification

Ph.D. Dissertation

Andreas Møgelmose

Dissertation submitted July 29, 2015

Thesis submitted: August 2015

PhD supervisor: Prof. Thomas B. Moeslund
Aalborg University

PhD committee: Associate Professor Claus B. Madsen
Aalborg University (committee chair)

Professor Dr. Dariu M. Gavrila
Daimler Research & Development
University of Amsterdam

Docent Henrik Karstoft
Aarhus University

PhD Series: Faculty of Engineering and Science, Aalborg University

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-333-3

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2015

Curriculum Vitae

Andreas Møgelmose

Andreas Møgelmose received his Bachelor's degree in Computer Engineering on the topic of Information Processing Systems in 2010 and his MSc in Informatics in 2012, both from Aalborg University, Denmark. In August 2012, he embarked on the PhD study with the Visual Analysis of People Lab at the section of Media Technology, Aalborg University.

Furthermore, Andreas is a part of the LISA lab at the University of California, San Diego in the Computer Vision and Robotics Research Laboratory, where he has been a visiting scholar multiple times.

His main interests are computer vision and machine learning, especially in the area of detecting people, pedestrians, and traffic signs and signals, as well as re-identification of people. He has been involved in supervision of undergraduate and graduate students within image processing and computer vision.

Abstract

Automated analysis of traffic situations - be it cars, signs, or pedestrians - is becoming increasingly relevant and feasible with the advent of powerful sensors, computers, and algorithms. This PhD thesis tackles three themes within this realm: Traffic sign detection, pedestrian detection and analysis, and person re-identification.

In traffic sign detection, the work comprises a thorough survey of the state of the art, assembly of the worlds largest public dataset with U.S. traffic signs, and work in machine learning based detection algorithms. It was shown that detection of U.S. traffic signs has traditionally lacked behind detection of European signs, which led to the effort of collecting the dataset and pushing the state of the art in detection performance for these signs by using the Aggregate Channel Features detector.

Within pedestrian detection, a method combining Viola-Jones and HOG/SVM has been put forth, which gives the speed advantage of Viola-Jones and the detection performance of HOG/SVM. Work has also been done in tracking the gaze of drivers to determine which pedestrians a driver may have missed. Finally, pedestrian tracking has been performed in an attempt to predict their future behavior in order to avoid dangerous situations.

Person re-identification has been attempted in a multi-modal fashion. Traditionally, re-identification has been performed using only RGB input from regular surveillance cameras, but we added depth and thermal information to the mix. Several iterations of a multi-modal system were tested, but the advantage of the additional information turned out to be limited.

Resumé

Automatisk analyse af trafiksituationer - for eksempel biler, skilte eller fodgængere - bliver mere og mere relevant i takt med at der udvikles bedre sensorer, computere og algoritmer. Denne PhD-afhandling behandler tre emner i denne verden: Detektion af skilte, detektion og analyse af fodgængere, samt person re-identifikation.

Indenfor skiltedetektion består arbejdet af en oversigtsartikel over forskningens nuværende niveau, indsamling af verdens største datasæt med billeder af amerikanske skilte, og arbejde med machine learning-baserede detektionsalgoritmer. Det fremgår at detektion af amerikanske skilte traditionelt har haltet bagefter detektion af europæiske skilte, hvilket førte til indsamling af datasættet og senere til fremskridt i detektionen af disse skilte ved hjælp af Aggregate Channel Features metoden.

I fodgængerdetektion præsenteres en metode der kombinerer Viola-Jones og HOG/SVM. Den giver hastighedsfordelen fra Viola-Jones, men med detektionssikkerheden fra HOG/SVM. Der er også blevet arbejdet med at aflæse synsretningen for chauffører for at fastslå hvilke fodgængere en chauffør eventuelt har overset. Endelig er der udført tracking af fodgængere i et forsøg på at forudsige deres fremtidige opførsel, så farlige situationer kan undgås før de opstår.

Re-identifikation af personer er blever forsøgt med multimodal data. Traditionelt udføres det blot med RGB input fra almindelige overvågningskameraer, men vi har tilføjet dybdedata og termisk data. Flere udgaver at et multimodal re-identifikationssystem blev testet, men fordelen ved at tilføje de ekstra datatyper har vist sig at være stærkt begrænset.

Contents

Thesis Details

Thesis Title: Visual Analysis in Traffic & Re-identification
Ph.D. Student: Andreas Møgelmose
Supervisor: Professor Thomas B. Moeslund, Aalborg University

The main body of this thesis consists of the following papers:

Traffic Sign Detection

[A] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey". In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (Dec. 2012), pp. 1484–1497

[B] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets". In: *21st International Conference on Pattern Recognition (ICPR)*. IEEE, 2012, pp. 3452–3455

[C] Andreas Møgelmose, Dongran Liu, and Mohan M. Trivedi. "Detection of US Traffic Signs". In: *IEEE Transactions on Intelligent Transportation Systems* In press (2015)

Pedestrian Detection and Analysis

[D] Antonio Prioletti, Andreas Møgelmose, Paolo Grisleri, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund. "Part-Based Pedestrian Detection and Feature-Based Tracking for Driver Assistance: Real-Time, Robust Algorithms, and Evaluation." In: *IEEE Transactions on Intelligent Transportation Systems* 14.3 (Sept. 2013), pp. 1346–1359

[E] Ashish Tawari, Andreas Møgelmose, Sujitha Martin, Thomas B. Moeslund, and Mohan M. Trivedi. "Attention Estimation by Simultaneous

Analysis of Viewer and View". In: *17th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2014, pp. 1381–1387

[F] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Trajectory Analysis and Prediction for improved Pedestrian Safety: Integrated Framework and Evaluations". In: *Intelligent Vehicles Symposium (IV)*. In press. IEEE, June 2015

Person Re-Identification

[G] Andreas Møgelmose, Thomas B. Moeslund, and Kamal Nasrollahi. "Multimodal Person Re-Identification using RGB-D Sensors and a Transient Identification Database". In: *International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2013, pp. 1–4

[H] Andreas Møgelmose, Chris Bahnsen, and Thomas B. Moeslund. "Comparison of Multi-shot Models for Short-term Re-identification of People using RGB-D Sensors". In: *11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP2015)*. Mar. 2015

[I] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund, and Sergio Escalera. "Tri-modal Person Re-identification with RGB, Depth and Thermal Features". In: *9th Workshop on Perception Beyond the Visible Spectrum, CVPR Workshops*. IEEE, 2013, pp. 301–307

In addition to the main papers listed above, the following publications have been co-authored in connection with the PhD:

- Andreas Møgelmose, Antonio Prioletti, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund. "Two-Stage Part-Based Pedestrian Detection". In: *15th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sept. 2012, pp. 73–77
- Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Traffic Sign Detection and Analysis: Recent Studies and Emerging Trends". In: *15th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sept. 2012, pp. 1310–1314
- Andreas Møgelmose, Dongran Liu, and Mohan M. Trivedi. "Traffic Sign Detection for U.S. Roads: Remaining Challenges and a Case for Tracking". In: *17th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2014, pp. 1394–1399

- Sujitha Martin, Eshed Ohn-Bar, Ashish Tawari, Andreas Møgelmose, and Mohan M. Trivedi. "Vision for Intelligent Vehicles and Applications: A Challenging in-the Wild Dataset". In: *The Future of Datasets in Vision, CVPR Workshops*. IEEE, June 2015

- Mark P. Philipsen, Morten B. Jensen, Andreas Møgelmose, Thomas B. Moeslund, and Mohan M. Trivedi. "Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives". In: *IEEE Transactions on Intelligent Transportation Systems* In press (2015)

- Mark P. Philipsen, Morten B. Jensen, Andreas Møgelmose, Thomas B. Moeslund, and Mohan M. Trivedi. "Learning Based Traffic Light Detection: Evaluation on Challenging Dataset". In: *18th Intelligent Transportation Systems Conference (ITSC)*. Submitted. IEEE. IEEE, 2015

- Mark P. Philipsen, Morten B. Jensen, Ravi K. Satzoda, Mohan M. Trivedi, Andreas Møgelmose, and Thomas B. Moeslund. "Night-Time Drive Analysis using Stereo Vision for Data Reduction in Naturalistic Driving Studies". In: *Intelligent Vehicles Symposium (IV)*. IEEE, June 2015

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

Preface

This thesis is submitted as a collection of papers in partial fulfillment of a PhD study at the Section of Media Technology, Aalborg University, Denmark. The work consists of three parts: Traffic Sign Detection, Pedestrian Detection and Analysis, and Person Re-identification. Thus, this thesis is organized with four parts. Part one contains an overview of the state-of-the-art in these three fields and the contributions which have been made to them during this work. This is followed by one part containing selected papers published as part of this PhD in each of the three fields.

This project has been carried out from 2012-2015, mainly in the Visual Analysis of People Lab at Aalborg University, but with two research stays in Laboratory of Intelligent and Safe Automobiles (LISA) at University of California, San Diego.

I would like to thank my supervisor prof. Thomas B. Moeslund for good supervision throughout my master's and PhD, as well as prof. Mohan M. Trivedi for his always enthusiastic supervision while I was in San Diego. A thank goes out to my colleagues and friends in the two labs, especially Anders Jørgensen, Chris Bahnsen, and Rikke Gade in Aalborg and Eshed Ohn-Bar, Sujitha Martin, Ashish Tawari, and Ravi Satzoda in San Diego. Finally, I would like to thank my good friends and fellow nerds Anders Tornvig, Joachim Kristensen, and Johan Ejstrud for listening to all my rants about PhD'ing, computer vision, world travel, and everything in between. And for having a beer with me and laughing at it all, when things did not always work as I wanted them too.

Andreas Møgelmose
Aalborg University, July 29, 2015

Part I

Overview of the work

Chapter 1

Introduction

As sensors and computers have progressed over the past 40 years, computer vision has emerged as one of the most interesting and diverse fields in modern technology. While still a relatively young field, computer vision is showing great promise in making computers understand the world around them - a skill which is essential in many automation tasks. Computer vision makes robots able to see and reason about their environment. It also automates many surveillance tasks, freeing people from the tedious task of watching endless hours of eventless video.

This thesis tackles two seemingly small corners of computer vision: computer vision in traffic, and computer vision in re-identification of people in relation to surveillance. And as with anything else you investigate in sufficient detail, it turns out that these corners might not be so small after all.

1 Traffic analysis

Traffic was changed forever at the turn of the 19th century, as automobiles began to be widely available. Transportation became faster, easier, and cheaper as years passed. But along with the upsides came new risks for people using cars and people around cars. In the very early years, cars were required to have a man with a bell walking in front, warning other road users about it. That is no wonder, because as cars invaded large cities with insufficient infrastructure, accident rates soared. Even today, a hundred years later, cars are still the main cause of injury-related deaths worldwide[2].

Traffic accidents can have many causes, but driver error is the predominant one, accounting for more than 90% of accidents[3, 1]. Thus, in order to reduce traffic accidents and related injuries, it is imperative to reduce the number of driver errors - of course along with research in making cars safe

Fig. 1.1: Detecting pedestrians is an essential part of traffic analysis.

when an accident inevitably happens. Reducing driver errors can take two primary avenues: Remove the driver from the equation completely by making the car fully autonomous, or help the driver control the car in a safe way. At this point, autonomous cars seem a certainty, probably after a transition period with increasingly capable driver assistance systems. In either case, computer vision is an integral part of the solution.

In order for cars to drive themselves, or help humans drive, they need to gain some form of understanding of the world around them. They need to detect driveable surfaces, see obstacles, and read signals and signs. Many obstacles can be found using range sensors such as lidars and radars, some context can be gained by looking at detailed maps, but other kinds rely on a sense of sight.

This thesis presents works related to detection of traffic signs, as well as detection and understanding of pedestrians in traffic.

2 Re-identification

Another somewhat unrelated research area, which is also treated in this thesis, is that of person re-identification.

In many large areas and structures, it is beneficial to understand how people move around. In airports, the management is interested in eliminating bottlenecks around security checks and corridors, and in routing people through the building in an efficient manner (or through a lot of shops in a

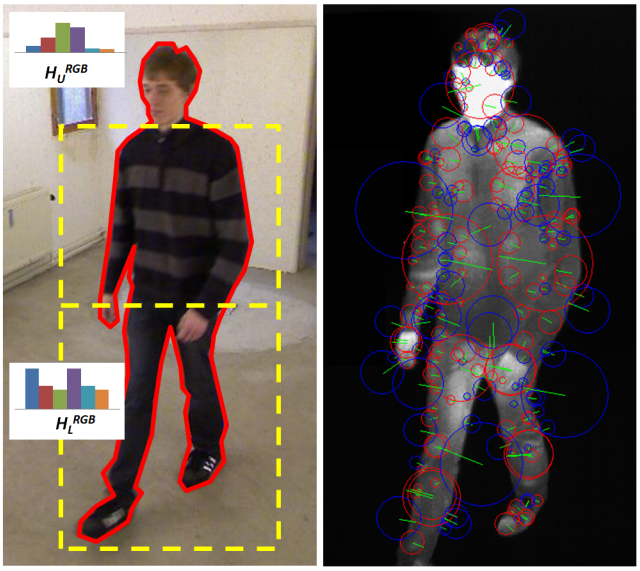


Fig. 1.2: Re-identification is about capturing distinctive features from each of the passing persons - in this case RGB histograms and SURF-features.

profitable manner). In amusement parks, queue times are interesting, and in malls, movement patterns can help optimize the design and layout of stores.

One way to gather this information is to track specific people - not necessarily everybody - through camera surveillance. In some cases one can rely on a surveillance network with full coverage, but often the fields of view are not overlapping between cameras. This means that in order to track people, it is necessary to recognize them by looks.

That is what re-identification is about: Finding distinct characteristics of observed people in order to recognize them later. Re-identification should not require the participation of the subjects, and it should work at a distance with low-resolution surveillance cameras. This means that traditional James-Bond-esque face recogtion methods fall short. Instead we rely on soft biometrics, such as clothing and hair color.

This thesis looks at re-identification in a multi-modal framework, trying to leverage depth and thermal view on top of traditional color video.

While the topics of traffic and re-identification are not immediately related, there is a certain overlap in methods, especially when it comes to detection, which is necessary in both applications. The data quality is also generally low for both, which presents some significant shared challenges.

3 Thesis structure

This thesis is divided into 4 parts. Part I contains an introduction to each of the three main topics: *Traffic Sign Detection*, *Pedestrian Detection and Analysis*, and *Person Re-identification*. Each of these chapters also presents an overview of the current state-of-the-art of that topic, along with a description of the contributions from this PhD work. Part I is wrapped up with a general conclusion. This first part is meant as a general introduction and will as such not delve deeply into the details of the work in this PhD. For that, the reader is referred to the enclosed articles, which are also referenced throughout the text when relevant.

After the introduction, parts II through IV contain the included papers for each of the three topics. Several additional papers have been written during the PhD, and while these are not included in the thesis, they will be referenced whenever appropriate. Each chapter contains its own bibliography.

Bibliography

[1] *National Motor Vehicle Crash Causation Survey*. Tech. rep. NHTSA, 2008.

[2] M. Peden, R. Scurfield, and D. Sleet. *World report on road traffic injury prevention*. Tech. rep. World Health Organization, 2004.

[3] J. R. Treat, N. S. Tumbas, S. T. McDonald, D. Shinar, R. D. Hume, R. E. Mayer, R. L. Stansifer, and N. J. Castellan. *Tri-level study of the causes of traffic accidents: final report. Executive summary.* Tech. rep. NHTSA, 1979.

Chapter 2

Traffic Sign Detection

This chapter provides an overview of the problem of Traffic Sign Detection (TSD) and its current state-of-the-art. Inevitably, it will also touch upon the related problem of Traffic Sign Classification, which combined with the process of detection is called Traffic Sign Recognition (TSR). TSR is relevant both in relation to self-driving cars and driver assistance systems, but the core methods are the same for either application, so this chapter is not going to distinguish between these applications. Much of the text in this chapter is taken from [18] and [14], but edited and abbreviated.

1 Introduction

Automated Traffic Sign Recognition is about reading the traffic signs posted along the road on which a car is driving. As mentioned above, this has multiple applications [18]:

1. Driver assistance systems: Assist the driver by informing of current restrictions, limits, and warnings.
2. Intelligent autonomous vehicles: Any autonomous car that is to drive on public roads must have a means of obtaining the current traffic regulations. This can be done through TSR.
3. Highway maintenance: Check the presence and condition of signs along major roads.
4. Sign inventory: Similar to the above task, create an inventory of signs in city environments.

In the early days (pre-2010) of TSR research, the inventory application was often cited [5, 10, 11], as this did not necessarily put any real-time performance restrictions on the developed systems. More recently, the focus has

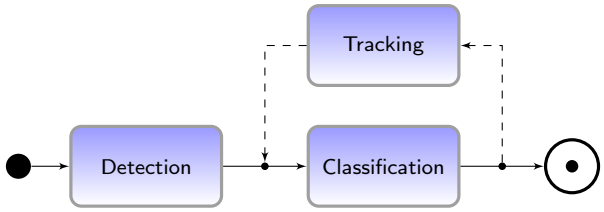


Fig. 2.1: The basic flow in most TSR systems.

moved to real-time TSR systems, as car manufacturers proceed to develop speed limit sign detection (most commonly) for use in their cars, and as universities and car manufacturers alike research autonomous cars.

As mentioned, TSR is generally split into two stages: Detection and classification (see fig. 2.1). Detection is concerned with locating signs in input images, while classification is about determining what type of sign the system is looking at. The two tasks can often be treated as completely separate, but in some cases the classifier relies on the detector to supply information, such as the sign shape or sign size. In a full system, the two stages are depending on each other and it does not make sense to have a classifier without a detection stage. This chapter focuses primarily on detection.

Many TSR systems are tailored to a specific sign type. Due to the vast differences in sign design from region to region (see the following section), and the differences in sign design based on their purpose, many systems narrow their scope down to a specific sign type in a specific country.

1.1 Traffic signs

Traffic signs are markers placed along roads to inform drivers about either road conditions and restrictions or which direction to go. They communicate a wealth of information, but are designed to do so efficiently and at a glance. This also means that they are often designed to stand out from their surroundings, making the detection task fairly well defined.

The designs of traffic signs are standardized through laws, but differ across the world. In Europe many signs are standardized via the Vienna Convention on Road Signs And Signals [27]. There, shapes are used to categorize different types of signs: Circular signs are prohibitions including speed limits, triangular signs are warnings and rectangular signs are used for recommendations or sub-signs in conjunction with one of the other shapes. In addition to these, octagonal signs are used to signal a full stop, downwards pointing triangles yield, and countries have other different types, e.g. to inform about city limits. Examples of these signs can be seen in fig. 2.2.

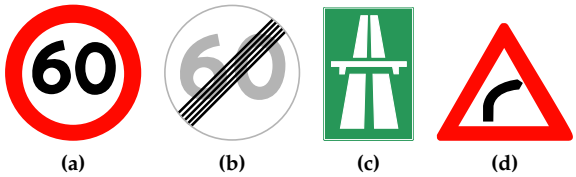


Fig. 2.2: Examples of European signs. These are Danish, but many countries use similar signs. (a) Speed limit. Sign C55. (b) End speed limit. Sign C56. (c) Start of freeway. Sign E55. (d) Right turn. Sign A41.

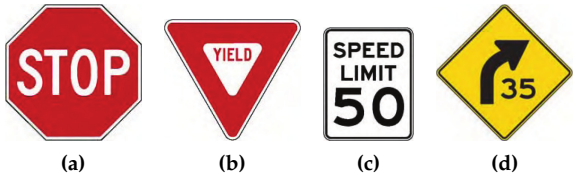


Fig. 2.3: Examples of signs from the US national MUTCD. (a) Stop. Sign R1-1. (b) Yield. Sign R1-2. (c) Speed limit. Sign R2-1. (d) Turn warning with speed recommendation. Sign W1-2a.. Image source: [24]

In the US, traffic signs are regulated by the Manual on Uniform Traffic Control Devices (MUTCD) [24]. It defines which signs exist and how they should be used. It is accompanied by the Standard Highway Signs and Markings (SHSM) book, which describes the exact designs and measurements of signs. At the time of writing, the most recent MUTCD was from 2009, and the SHSM book was updated in 2012. The MUTCD contains a few hundred different signs, divided into 13 categories.

To complicate matters further, each US state can decide whether it wishes to follow the MUTCD. A state has three options:

1. Adopt the MUTCD fully as is.
2. Adopt the MUTCD but add a State Supplement.
3. Adopt a State MUTCD that is "in substantial conformance with" the national MUTCD.

19 US states have adopted the national MUTCD without modifications, 23 have adopted the national MUTCD with a state supplement and 10 have opted to create a State MUTCD (the count includes the District of Columbia and Puerto Rico). Examples of US signs can be seen in fig. 2.3.

New Zealand uses a sign standard with warning signs that are yellow diamonds, as in the US, but regulatory signs that are round with a red bor-

(a) (b) (c)

Fig. 2.4: Examples of similar signs from the MUTCD. (a) Speed limit. Sign R2-1. (b) Minimum speed. Sign R2-4. (c) End speed limit. Sign R3 (CA), exists only in the California MUTCD. Image source: [24]

der, like the ones from the Vienna Convention countries. Japan uses signs that are generally in compliance with the Vienna Convention, as are Chinese regulatory signs. Chinese warning signs are triangular with a black/yellow color scheme. Central and South American countries do not participate in any international standard, but often use signs somewhat like the American standard.

The bulk of the research in TSR systems has laid in European signs [18], but the differences in sign designs matter very much in a detection context. The top row of fig. 2.5 shows typical signs from each of the major sign superclasses in Europe: Mandatory, danger, and prohibitory. Each class is very distinctive, not only from the others, but also from most things in the real world. They all have both a rather distinctive shape and a strongly colored border/background. US signs are not in exactly the same classes, but the bottom row shows the matching US signs. Fig. 2.6 shows more examples of US signs. From the outset, three things are obvious:

1. US signs bear little to no resemblance to Vienna Convention signs, at the very least requiring re-training of any detector.

2. The strong visual structure in Vienna Convention signs is less present in US signs. The strongest visual clue for US signs is the yellow diamond of warning signs, but even that is not present for all warning signs. The stop sign (which is identical with its Vienna Convention counterpart) is also visually strong. However, most other signs are just white rectangles of varying aspect ratios, which should be challenging to standard detectors which often rely heavily on color cues.

3. Many US signs contain only text to convey their message, as opposed to Vienna Convention signs which mostly use icons and pictograms.

Given these large differences to Vienna Convention signs, and the size of the car market in the US, it is surprising that in [18], only two studies were concerned with US traffic signs, and as we describe in the previous section on related studies, not many have come out since the publication of the review.

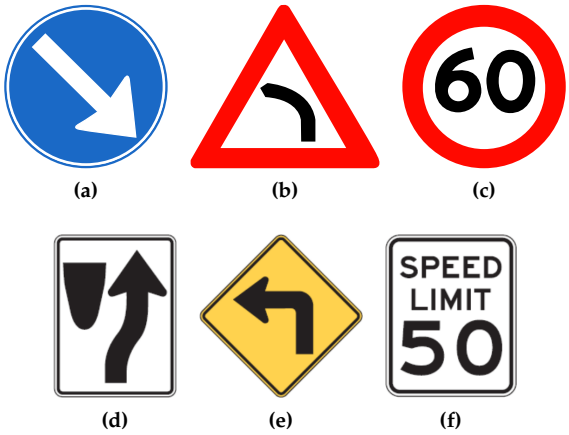


Fig. 2.5: Vienna convention signs (top row) along with their US counterparts (bottom row). (a) Keep right, superclass mandatory. Sign D15,3. (b) Left turn, superclass danger. Sign A41-2. (c) 60 km/h speed limit, superclass prohibitory. Sign C55. (d) Keep right, superclass traffic movement. Sign R4-7. (e) Left turn, superclass warning. Sign W1-1. (f) 50 mph speed limit, superclass speed limit. Sign R2-1. Image source (d)-(f): [24]

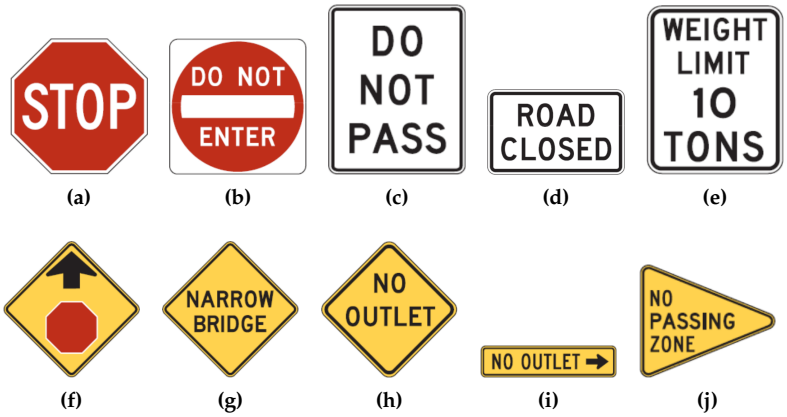


Fig. 2.6: Examples of US signs. (a) Sign R1-1. (b) Sign R5-1. (c) Sign R4-1. (d) Sign R1-2. (e) Sign R12-1. (f) Sign W3-1. (g) Sign W5-2. (h) Sign W14-2. (i) Sign W14-2a. (j) Sign W14-3. Image source: [24]

While signs are well defined through laws and designed to be easy to spot, there are still plenty of challenges for TSR systems. They include:

- Signs being similar within or across categories (see fig. 2.4).
- Signs may be faded or dirty, thus no longer their specified color.
- Bent sign posts, so the sign is no longer orthogonal to the road.
- Lighting conditions may make color detection unreliable.
- Low contrast may make shape detection hard.
- In cluttered urban environments, other objects may look very similar.
- Varying weather conditions.

2 State of the Art

Traffic sign detection has been researched seriously for about a decade. All the way up until 2013, contributions were very hard to compare, since most researches would test on their own private datasets[18], as opposed to public ones, which allow for cross-comparison between systems. In 2013, the German Traffic Sign Detection Benchmark (GTSDB)[6] was launched, which really kickstarted the research in machine learning based detectors and pushed the state-of-the-art detection performance to near-perfection on European signs.

Before the GTSDB, there was little use of machine learning in TSD, except for a few HOG/SVM based approaches[18]. The approaches could generally be divided into two groups: Color-based approaches and shape-based approaches, sometimes in various combinations, but generally using a manually defined model with hand-tunes parameters. Color based methods would take advantage of the fact that traffic signs are designed to be easily distinguished from their surroundings, often colored in highly visible contrasting colors. These colors are extracted from the input image and used as a base for the detection. Just like signs have specific colors, they also have very well defined shapes that can be searched for. Shape based methods ignore the color in favor of the characteristic shape of signs.

In either case, the detector would be finely tuned to a specific sign class, and it may not even be generalizable to arbitrary sign classes. For example, purely color-based methods work poorly with white signs like US speed limit signs. Since these methods are largely outdated now, this chapter will not describe them in any further depth. They are, however, described in detail in chapter A. One thing of note is that the vast majority of the early work was concerned with European signs only, as their designs lend themselves well to

color-segmentation.

Recently the learning-based methods have taken over completely in defining the state-of-the-art. Thus, all the front-runners in the GTSDB competition were learning based. The competition encompassed 18 teams and 3 teams were considered the top performers: *Team VISICS*[13], *Team Litsi*[8], and *Team wgy@HIT501*[28]. *Team VISICS* use the Integral Channel Features (known as *ChnFtrs* or ICF) proposed by Dollár et. al.[4] for pedestrian detection and later improved in [3]. *Team Litsi* first establishes regions of interest with color classification and shape matching and the detect signs using HOG and color histograms with an SVM, features somewhat similar to ICF. Finally, *Team wgy@HIT501* uses HOG features, finding candidates with LDA and performing a more fine-grained detection using HOG with IK-SVM. In essence, all three approaches are rather similar, especially when it comes to features. Another recent paper presenting work on the GTSDB dataset is [9], which shows somewhat worse detection performance than the competitors above, but at a faster speed.

For US traffic signs, the activity has been less enthusiastic. Only a few recent studies take on US traffic signs. In 2012, [16] (see chapter B) evaluated whether synthetic training data of US signs could be used successfully to train a rudimentary detector, but performance for the synthetic training data was poor compared to real-world images. Also in 2012, Staudenmaier et. al. [25] (building on their previous paper [26]) showed a Bayesian Classifier Cascade with intensity features and tensor features (which describe edges). They detect US speed limit signs at a good rate above 90%, but with several false positives per image, much, much worse than the current European state-of-the-art systems. Abukhait et. al. [1] use a shape-based detector to find US speed limit signs - note the model-based, rather than learning-based approach. The detector is part of a full recognition system, and the only reported performance figure is a detection rate of about 88%, but without mention of false positive rates. Stepping back in time to 2008, Keller et. al. [7] worked on detection of US speed limit signs using a voting based rectangle detector (see also [2]) followed by an AdaBoost classifier. Moutarde et. al. [19, 20] also tackled the case of US speed limit signs using a proprietary rectangle detector. In 2015, we published [14] (chapter C), which presents state-of-the-art performance on US signs using the ACF detector, comparable to the results seen in the GTSDB. Before this, US sign detection lacked severely behind European signs.

3 Contributions

The work done as part of this PhD has mainly aimed at pushing the state-of-the-art in detection of US traffic signs, as hinted above. The first major study published was *Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey* [18] (included in chapter A). It contained a comprehensive overview of the state-of-the-art in TSD - the first of its kind. Two main conclusions were reached:

1. Previous work had not used standardized datasets for testing, making comparison of systems extremely difficult.

2. There was almost no research done in the detection of US traffic signs, nor was any US traffic sign dataset available.

As a direct consequence of the findings, we assembled the world's first publicly available annotated dataset of US traffic signs, which was released along with the survey. At that time it contained 7855 annotations. An abbreviated version of the survey was also published at the International Conference of Intelligent Transportation Systems in 2012 [17].

The assembly of the dataset revealed - unsurprisingly - that it is a very time consuming job. This sparked interest in trying to use synthetic training data in TSD. Signs are very well defined, which makes them easy to synthesize. These investigations were presented in *Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets* [16] (included in chapter B). Synthetic training images were generated by applying a number of distortions to drawn versions of the signs. While these distorted templates were quite similar to real-world training data in a simple visual inspection, it turned out that they did not encompass all the distortions present in the real-world, and as such did not work well for training traffic sign detectors.

In 2014, we continued the quest of detecting US traffic signs, an area which is still not very well researched. We applied ICF, which had seen good success for detection of European signs, and as a first, we used it successor ACF in the domain of TSD. Initial results were published at at the International Conference of Intelligent Transportation Systems in 2014 [15], and the work later resulted in the journal paper *Detection of US Traffic Signs* [14] (included in chapter C). In this paper, we presented state-of-the-art detection rates on US traffic signs, and brought US TSD on par with European signs for several sign superclasses. US speed limit signs are still difficult to detect, and further research is needed.

In 2014 the public dataset was extended to 13493 annotations. The updated dataset was used in the VIVA Challenge[1] and was also used in [12].

[1] http://cvrr.ucsd.edu/vivachallenge/

Based on our work in the area of TSD, I co-supervised a group of master's students working on traffic light detection, and as part of that co-authored a survey [22] similar to the original one on traffic signs, a paper presenting an ACF-derived traffic light detection [21], and finally a paper on analyzing events in data captured in naturalistic driving studies [23].

Bibliography

[1] Jafar Abukhait, Imad Zyout, and Ayman M Mansour. "Speed Sign Recognition using Shape-based Features". In: *International Journal of Computer Applications* 84.15 (2013), pp. 31–37.

[2] N. Barnes and G. Loy. "Real-time regular polygonal sign detection". In: *Field and Service Robotics*. Springer. 2006, pp. 55–66.

[3] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. "Pedestrian detection at 100 frames per second". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2903–2910.

[4] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. "Integral Channel Features". In: *BMVC*. Vol. 2. 3. 2009, p. 5.

[5] M.A. Garcia-Garrido, M.A. Sotelo, and E. Martm-Gorostiza. "Fast traffic sign detection and recognition under changing lighting conditions". In: *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*. Sept. 2006, pp. 811–816.

[6] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark". In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE. 2013, pp. 1–8.

[7] C.G. Keller, C. Sprunk, C. Bahlmann, J. Giebel, and G. Baratoff. "Real-time recognition of U.S. speed signs". In: *Intelligent Vehicles Symposium, IEEE*. June 2008, pp. 518–523. DOI: 10.1109/IVS.2008.4621282.

[8] Ming Liang, Mingyi Yuan, Xiaolin Hu, Jianmin Li, and Huaping Liu. "Traffic sign detection by ROI extraction and histogram features-based recognition". In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. Aug. 2013, pp. 1–8. DOI: 10.1109/IJCNN.2013.6706810.

[9] Chunsheng Liu, Faliang Chang, and Zhenxue Chen. "Rapid Multiclass Traffic Sign Detection in High-Resolution Images". In: *Intelligent Transportation Systems, IEEE Transactions on* 15.6 (Dec. 2014), pp. 2394–2403. ISSN: 1524-9050. DOI: 10.1109/TITS.2014.2314711.

[10] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. López-Ferreras. "Road-sign detection and recognition based on support vector machines". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.2 (2007), pp. 264–278.

[11] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Siegmann, H. Gomez-Moreno, and F.J. Acevedo-Rodriguez. "Traffic sign recognition system for inventory purposes". In: *Intelligent Vehicles Symposium, 2008 IEEE*. June 2008, pp. 590–595. DOI: 10.1109/IVS.2008.4621233.

[12] Sujitha Martin, Eshed Ohn-Bar, Ashish Tawari, Andreas Møgelmose, and Mohan M. Trivedi. "Vision for Intelligent Vehicles and Applications: A Challenging in-the Wild Dataset". In: *The Future of Datasets in Vision, CVPR Workshops*. IEEE, June 2015.

[13] Markus Mathias, Radu Timofte, Rodrigo Benenson, and Luc Van Gool. "Traffic sign recognition — How far are we from the solution?" In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE. 2013, pp. 1–8.

[14] Andreas Møgelmose, Dongran Liu, and Mohan M. Trivedi. "Detection of US Traffic Signs". In: *IEEE Transactions on Intelligent Transportation Systems* In press (2015).

[15] Andreas Møgelmose, Dongran Liu, and Mohan M. Trivedi. "Traffic Sign Detection for U.S. Roads: Remaining Challenges and a Case for Tracking". In: *17th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2014, pp. 1394–1399.

[16] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets". In: *21st International Conference on Pattern Recognition (ICPR)*. IEEE, 2012, pp. 3452–3455.

[17] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Traffic Sign Detection and Analysis: Recent Studies and Emerging Trends". In: *15th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sept. 2012, pp. 1310–1314.

[18] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey". In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (Dec. 2012), pp. 1484–1497.

[19] F. Moutarde, A. Bargeton, A. Herbin, and L. Chanussot. "Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system". In: *Intelligent Vehicles Symposium*. IEEE. 2007, pp. 1122–1126.

[20] Fabien Moutarde, Alexandre Bargeton, Anne Herbin, and Lowik Chanussot. "Modular Traffic Sign Recognition applied to on-vehicle real-time visual detection of American and European speed limit signs". In: vol. 14. 2009.

[21] Mark P. Philipsen, Morten B. Jensen, Andreas Møgelmose, Thomas B. Moeslund, and Mohan M. Trivedi. "Learning Based Traffic Light Detection: Evaluation on Challenging Dataset". In: *18th Intelligent Transportation Systems Conference (ITSC)*. Submitted. IEEE. IEEE, 2015.

[22] Mark P. Philipsen, Morten B. Jensen, Andreas Møgelmose, Thomas B. Moeslund, and Mohan M. Trivedi. "Vision for Looking at Traffic Lights: Issues, Survey, and Perspectives". In: *IEEE Transactions on Intelligent Transportation Systems* In press (2015).

[23] Mark P. Philipsen, Morten B. Jensen, Ravi K. Satzoda, Mohan M. Trivedi, Andreas Møgelmose, and Thomas B. Moeslund. "Night-Time Drive Analysis using Stereo Vision for Data Reduction in Naturalistic Driving Studies". In: *Intelligent Vehicles Symposium (IV)*. IEEE, June 2015.

[24] State of California, Department of Transportation. *California Manual on Uniform Traffic Control Devices for Streets and Highways*.

[25] Armin Staudenmaier, Ulrich Klauck, Ulrich Kreßel, Frank Lindner, and Christian Wöhler. "Confidence Measurements for Adaptive Bayes Decision Classifier Cascades and Their Application to US Speed Limit Detection". In: *Pattern Recognition*. Vol. 7476. Lecture Notes in Computer Science. 2012, pp. 478–487.

[26] *Resource Optimized Cascaded Perceptron Classifiers using Structure Tensor Features for US Speed Limit Detection*. Vol. 12. 2011.

[27] United Nations Economic Commission for Europe. *Convention on Road Signs And Signals, of 1968*. 2006.

[28] Gangyi Wang, Guanghui Ren, Zhilu Wu, Yaqin Zhao, and Lihui Jiang. "A robust, coarse-to-fine traffic sign detection method". In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. Aug. 2013, pp. 1–5. DOI: 10.1109/IJCNN.2013.6706812.

Chapter 3

Pedestrian Detection and Analysis

In line with the theme of making intelligent cars understand their surroundings, this chapter provides an overview of the problem of pedestrian detection and analysis. Detection is obviously the task of finding pedestrians in images, whereas analysis in this case is about tracking them and predicting what they will do later. Pedestrian detection is somewhat related to traffic sign detection - or rather, traffic sign detection is somewhat related to pedestrian detection. Almost all recent traffic sign detection algorithms have their genesis in the pedestrian detection world. So not only is the problem highly relevant on its own, it also forms the foundation for detection in many other scenarios.

1 Introduction

Pedestrian detection and analysis is a relatively wide area, which has seen large amounts of research recently. As mentioned, we split the task into two individual subtasks: Detection and analysis. Detection has seen the most work and has been an active field almost since the beginning of computer vision, whereas analysis has gained traction over the past 3-5 years as detection has gotten more robust. In the following two sections, the two problems are introduced.

1.1 Pedestrian detection

The objective of pedestrian detection is simple: To find out if and where pedestrians are present in an input image. It is relevant in any context where

mobile robots move around people and in surveillance, but the most obvious application is in autonomous cars and driver assistance systems.

Cars navigating the world need a way to avoid hitting obstacles, especially people. Radars and lidars can provide the most basic obstacle detection, but have difficulty going any further than "there is something here." With camera-based detection we can know not just if an obstacle is present, but also that it is a person, and can take appropriate precautions based on their context.

Detecting a pedestrian is generally defined as drawing a bounding box around them. The specific act of detection is not concerned with their pose, looks, activity, or even position in the 3D space. The result is purely a 2D bounding box on the input image.

Pedestrian detection is not easy, as outlined in [14]:

- High variability in appearance among pedestrians: Pedestrians are physically very different. Not only are children much smaller than adults, even among adults there is a vast variation of body-types, clothing and postures. Futhermore, people move around and their potentially highly articulated poses make creating a unified detector very difficult.
- Cluttered backgrounds: In some surveillance contexts the background might be uniform and clearly separated from the foreground, but in cities windows, cars, chimneys, traffic signs, trees, and everything else in the world conspire to make very complex and cluttered scenes where the background is not easily discernible. There will also virtually always be many different layers in a scene and there is no guarantee that pedestrians are all at the same distance from the observer.
- Highly dynamic scenes with both pedestrian and camera motion: When talking driver assistance systems, the observed scene is constantly changing. Pedestrians move around, are intermittently occluded by cars, enter and exit houses. To complicate matters further, the camera on the ego-vehicle is also moving, constantly changing the background.
- Strict requirements in both speed and reliability: If these systems are to work with moving vehicles, they must never (for some definition of never) miss a pedestrian, and they must be able to refresh their view of the world often enough to have an up-to-date picture of risks and challenges.

It is no surprise then, that even though pedestrian detection has been researched for years with great strides made, the perfect pedestrian detector does not yet exist.

(a)

(b)

Fig. 3.1: Examples of input frames for pedestrian detection systems. The difficulty and variation across scenarios is evident. (a) Image from the large public Caltech Pedestrian Dataset. (b) Image from the data captured for [22] (chapter F). In these examples, the pedestrians were manually annotated.

1.2 Pedestrian analysis

Pedestrian analysis has a wider scope than detection and does not have a singular definition. In essence, pedestrian analysis is anything a system computes after the pedestrian has been detected. To narrow it down a bit, this chapter is about intent prediction. If the car can predict what a pedestrian is about to do next, it stands a much better chance of determining the proper course of action, than if it only knows the position of the person. Note that the prediction horizon is most often (but not always) relatively short, measured in fractions of a second into the future. Humans are very unpredictable from a machine's point of view, so projecting their detailed actions many seconds or minutes into the future is still not feasible.

As we shall see, pedestrian analysis does not have the same long history as pedestrian detection does. Pedestrian analysis hinges on the ability to detect pedestrians reliably, and thus, the research has taken a back seat to building a good detector. As detectors have become better over the past few years, pedestrian analysis has slowly emerged as a field of its own.

Pedestrian analysis can roughly be divided into three approaches, though they are often combined:

1. Tracking-based analysis
2. Motion cue analysis
3. Environment path analysis

Tracking is about building and understanding a kinematic model of the pedestrian. This means analyzing her motion over time and using that information to predict where she will be next. Highly simplified: If a person is running fast in one direction, it is likely that she will be further along her path in the following time instance, rather than suddenly moving backwards. A runner can change direction of course, but it is important to understand that the prediction necessary is only fractions of a second up to a second into the future. So while tracking is not necessarily the holy grail of pedestrian analysis, it provides a robust rough framework in which to reason about the intention of moving objects. It also puts no real requirements on the image quality, as long as pedestrians are detected: no more information than their position over time is necessary.

Motion cue analysis ignores these macro movements in preference of smaller detail movements. Instead of looking at the speed of the body's center of gravity, it looks at motions of arms and legs. Imagine a pedestrian waiting to cross the street. The speed of the person is zero, so the tracking approach will simply predict the subject to be still. The motion cue approach

attempts to detect as soon as the subject lifts her legs or arm, beginning to take a step. It can also try to estimate the orientation of the pedestrian's upper body and use that as a cue to where the person will go next. Motion cue analysis is interesting because it gives the ability to detect motion before the person has gained any speed, but it puts more requirements on the data quality than simple tracking: the pictures must have a resolution which allows for robust detection of body part motion.

Environment path analysis looks at the current environment and tries to infer scene context: roads, buildings, sidewalks, etc. By learning from how people move around in these environments, the system tries to predict how a particular pedestrian will behave given the circumstances of the scene. In some cases, tracking of pedestrians are utilized to refine the estimate, as some paths become increasingly likely as time progresses. This type of analysis has the advantage of allowing predictions into the far future (several seconds), but is most often less useful in predicting the immediate next position.

2 State of the Art

As the two sub-fields of this chapter are somewhat separate, each has its own state of the art section below.

2.1 Pedestrian detection

Pedestrian detection really took off with the advent of the Viola-Jones method [28] in 2001. It was first applied to face detection, but quickly adapted to pedestrians [29]. It learns Haar-wavelet-like features in a boosted cascade, which is a complicated way of saying that it looks at differences in brightness at specific places. As an example for faces: The area around the eyes is usually darker than the nose and cheeks, because the eyes sit in the shade of the forehead (see fig. 3.2). The method then uses a lot of these comparisons to filter away candidates which are not the searched for object. A clever trick, introduced with this method, is the integral image, which provides a fast and efficient way to sum pixel values under a given rectangle.

The Viola-Jones method - also known as boosted Haar-cascade - was at the forefront for several years, but was not great at capturing edge information. In 2005, Dalal and Triggs introduced HOG-SVM [3], which was aimed at pedestrian detection from the beginning. The main contribution of that work was the Histogram of Oriented Gradients - HOG - features, which quantifies edge strength and orientation in an image patch. Distinctive shapes, such as the head-shoulder outline of a person, also result in distinctive HOG-features. Dalal and triggs computes these features densely over a search window and

Fig. 3.2: Haar-like feature overlaid on a face. The value of this feature is computed by summing the values of the pixels under the black rectangle and subtracting the sum of the pixels under the white rectangle. Image credit: Soumyanilcsc / CC-BY-SA-4.0 / http://enwp.org?curid=37353395

Fig. 3.3: Example pictures from the INRIA pedestrian dataset. Image source: [3]

uses an SVM-classifier to determine whether it contains a pedestrian or not. HOG-SVM performed a lot better than Viola-Jones, in fact so much better that Dalal and Triggs had to compile a new dataset with more difficult-to-detect pedestrians than had previously been used. While previous detectors had been tested on the MIT dataset, the benchmark was now the significantly more challenging INRIA dataset, fig. 3.3.

If Viola-Jones signified the take-off of pedestrian detection, then HOG-SVM marked the beginning of the jet age. In the following years a number of new methods and datasets emerged. Of particular note is the Deformable Parts Model (DPM) and the Integral Channel Feature (ICF).

DPM, championed by Felzenszwalb et. al. [11], came out in 2008. It was based on the fact that people are highly articulated, so instead of looking at full bodies, a detector should instead find individual body parts, make sure they were in a sensible spatial configuration, and call that a detection. Vanilla DPM still uses the HOG-SVM detector to find the body parts. This idea saw a lot of research in the following years, but there is still no clear evidence as to the benefits of this approach[1].

ICF was put forth by Dollár et. al. in 2009[6]. It uses parts of the methods from both Viola-Jones and HOG. Specifically, it generalizes the computation of Haar-like features on integral images to an arbitrary set of "channels," fig. 3.4. A channel is a transformation of the input image, and can be a color

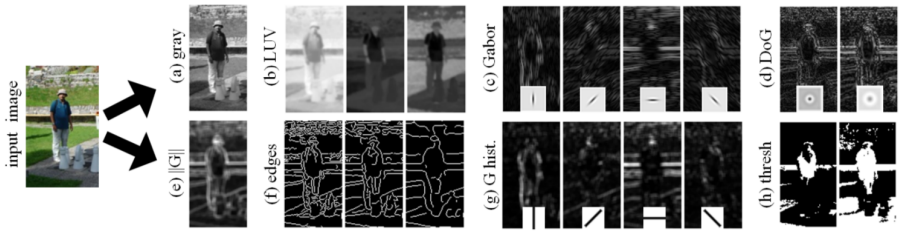


Fig. 3.4: A visual representation of some possible channels in ICF. Image source: [5]

channel, or - to emulate HOG - a set of edges with a specific orientation. The standard ICF implementation uses 10 different channels describing colors and edges. The ICF features are then classified as a pedestrian or not using a boosted decision forest. With ICF and its derivatives (Aggregate Channel Features, ACF, from 2014 is the latest iteration [4]), a very strong detector had been presented, which have only recently begun to be challenged by deep neural network based solutions.

Apart from the great influx of new detection algorithms, a wealth of new public pedestrian datasets also emerged in the wave of Dalal and Triggs' seminal paper. Soon, the INRIA dataset was outdated, and very large datasets, such as Daimler [9], KITTI [13] and Caltech-USA [5]. All of these have been instrumental in reaching the current level in pedestrian detection.

For further reading, "Pedestrian Detection: An Evaluation of the State of the Art" [7] and especially the excellent "Ten Years of Pedestrian Detection, What Have We Learned" [1] can be recommended.

2.2 Pedestrian analysis

With increasingly reliable detectors, parts of the research began to focus on analysing the actions of pedestrians, particularly prediction of pedestrians' future positions. As mentioned there are three approaches in this regards: Tracking, motion cue analysis, and environment path analysis. In this section we look at several examples of each approach, but it seems reasonable that a working system should incorporate predictions from all three. In contrast to the topics described previously in this text, there is no common benchmark and thus not really a way to determine the best method. Instead, we will investigate several representative systems below.

Tracking of moving pedestrians was done from a surveillance perspective in [23, 2], though the surveillance perspective means the a top-down view is used, making the tracking job easier. Closer to the subject of intelligent cars,

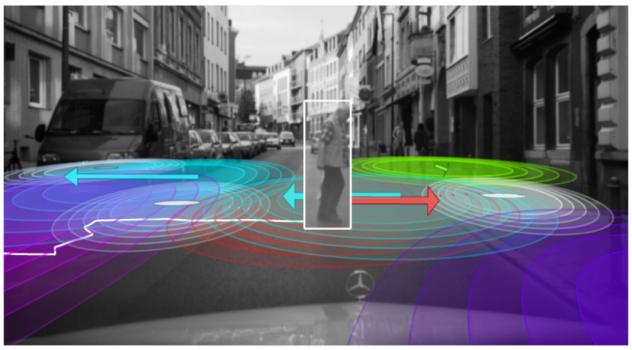


Fig. 3.5: Different avalable models for pedestrian behavior, which can be selected by analyzing the current pedestrian. Image source: [15]

tracking and prediction of pedestrians is done on stereo data in [25, 15] using Interacting Multiple Model Kalman Filters and SLDS tracking, respectively. Particularly interesting, in [15], multiple different models are built (see fig. 3.5), depending where the pedestrians are. When a new pedestrian enters view, a fitting model can be selected, which should provide a more reliable estimate.

In motion cue analysis, the papers [8, 12, 26] look at pedestrian orientation using different kinds of classifiers on static monocular pedestrian images. While none of them tackle prediction as such, a realiable classification of the body orientation is very relevant in this task. [19, 20, 10] do the same, but based on RGB-D data, and [10] even determines the orientation of the head and the torso separately. This is particularly interesting, as this information in combination should give a better indication of the pedestrian's intentions. In [18], local motion featues dubbed MCHOG are used to predict whether or not the pedestrian is about to take a step. The input data, however, is not coming from a car perspective, but a stationary multi-camera setup mounted at an intersection. A similar task is carried out in the very interesting [16], which uses optical flow and stereo data. This time on data from a real car. Though this is done in non-crowded scenarios with just a single pedestrian, it is one of the most advanced and capable pedestrian prediction systems at present.

The final prediction option is environment path analysis. This is most often done statically by analyzing image content, looking for path, with limited tracking information. In [17], this is exactly what is done. See fig. 3.6. An input image is automatically semantically labelled in superpixels, and based on these, the likely path is predictid. A very similar thing is done in [31].

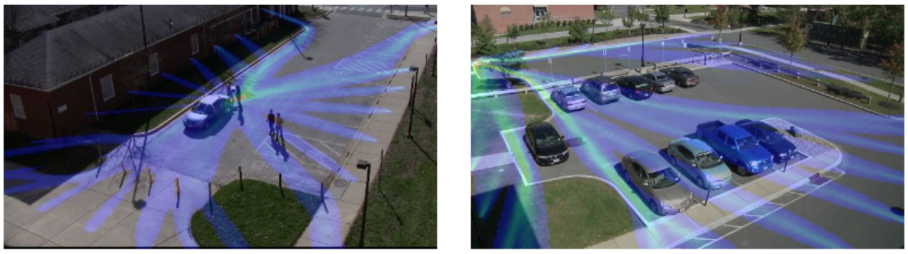


Fig. 3.6: Predicting plausible future paths by analyzing scene content. Image source: [17]

In [30], videos are analyzed to learn common motion patterns. New agents entering the scene is then matched to the most likely trajectory.

3 Contributions

During this PhD, a collection of work was done on pedestrian-related work. With regards to detection, we combined Viola-Jones and HOG-SVM in [24] (and a shorter conference version [21]). Viola-Jones was setup in a very permissive configuration, allowing many false positives through. All potential detections from the Viola-Jones stage was then passed on to the HOG stage, which tried to verify each. This way, Viola-Jones could act to find regions-of-interest for HOG. A major drawback of HOG-SVM is speed when a picture is densely searched, and this system speeded up detection significantly. The short-coming is of course that if a pedestrian is not found by Viola-Jones, she would not be found by HOG either. With appropriate tuning of parameters, this turned out not to be an issue. Apart from this detector combination, we also introduced doing detection in a coarse part-based manner, looking for upper and lower body, as well as full body. This could potentially help with occlusions. For further reading, see chapter D.

Within pedestrian prediction, we developed a particle filter based tracker working on monocular image data [22] (included in chapter F). Most of the previous work works with either a top-down view or RGB-D data, and we attempted to track and predict pedestrian behavior using only 2D input. This was combined with a mapping module, where a map was used to determine where the road is present in front of the car, and thus is a hazardous area for pedestrians. This way the driver can be warned if the system predicts a pedestrian to step onto the road.

Finally, we tried combining pedestrian detection with driver gaze direction estimation [27] (included in chapter E). The idea is that the driver should not be warned about pedestrians he has already seen, as we can assume he takes sufficient action to not hit those. Instead we attempted to detect the

pedestrian the driver had missed. We did this by detecting pedestrians in first-person video while tracking the gaze of the driver. These observations were correlated and the system was able to tell whether any given (detected) pedestrian had been noticed by the driver. The principle behind this system could be extended to traffic signs, traffic lights, other cars, or anything else we can detect and that the driver should be informed about. A system like this can thus cut down on information overload for the driver.

Bibliography

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. "Ten years of pedestrian detection, what have we learned?" In: *Computer Vision for Road Scene Understanding and Autonomous Driving (CVRSUAD, ECCV workshop)*. Sept. 2014.

[2] Aniket Bera, Nico Galoppo, Dillon Sharlet, Adam Lake, and Dinesh Manocha. "Adapt: real-time adaptive pedestrian tracking for crowded scenes". In: *Proceedings of Conference on Robotics and Automation, Hong Kong*. 2014.

[3] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*. 2005.

[4] P. Dollar, R. Appel, S. Belongie, and P. Perona. "Fast Feature Pyramids for Object Detection". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.8 (Aug. 2014), pp. 1532–1545. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2300479.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. "Pedestrian detection: A benchmark". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. June 2009, pp. 304–311. DOI: 10.1109/CVPR.2009.5206631.

[6] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. "Integral Channel Features". In: *BMVC*. Vol. 2. 3. 2009, p. 5.

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. "Pedestrian Detection: An Evaluation of the State of the Art". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.4 (2012), pp. 743–761.

[8] M. Enzweiler and D.M. Gavrila. "Integrated pedestrian classification and orientation estimation". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. June 2010, pp. 982–989. DOI: 10.1109/CVPR.2010.5540110.

[9] M. Enzweiler and D.M. Gavrila. "Monocular pedestrian detection: Survey and experiments". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.12 (2009), pp. 2179–2195.

[10] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila. "Joint probabilistic pedestrian head and body orientation estimation". In: *Intelligent Vehicles Symposium (IV), 2014 IEEE*. June 2014, pp. 617–622.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. "A discriminatively trained, multiscale, deformable part model". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (June 2008), pp. 1–8. ISSN: 1063-6919. DOI: 10.1109/CVPR.2008.4587597.

[12] T. Gandhi and Mohan M. Trivedi. "Image based estimation of pedestrian orientation for improving path prediction". In: *Intelligent Vehicles Symposium, 2008 IEEE*. June 2008, pp. 506–511. DOI: 10.1109/IVS.2008.4621257.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[14] D. Geronimo, A.M. Lopez, A.D. Sappa, and T. Graf. "Survey of pedestrian detection for advanced driver assistance systems". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.7 (2010), pp. 1239–1258.

[15] J. F. P. Kooij, N. Schneider, and D. M. Gavrila. "Analysis of pedestrian dynamics from a vehicle perspective". In: *Intelligent Vehicles Symposium (IV), 2014 IEEE*. June 2014, pp. 1445–1450.

[16] C.G. Keller and D.M. Gavrila. "Will the Pedestrian Cross? A Study on Pedestrian Path Prediction". In: *Intelligent Transportation Systems, IEEE Transactions on* 15.2 (Apr. 2014), pp. 494–506.

[17] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. "Activity forecasting". In: *Computer Vision–ECCV 2012*. Springer, 2012, pp. 201–214.

[18] S. Köhler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer. "Stationary Detection of the Pedestrian's Intention at Intersections". In: *Intelligent Transportation Systems Magazine, IEEE* 5.4 (winter 2013), pp. 87–99. ISSN: 1939-1390. DOI: 10.1109/MITS.2013.2276939.

[19] M. C. Liem and D. M. Gavrila. "Person appearance modeling and orientation estimation using Spherical Harmonics". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2013), pp. 1–6.

[20] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li. "Accurate Estimation of Human Body Orientation From RGB-D Sensors". In: *Cybernetics, IEEE Transactions on* 43.5 (Oct. 2013), pp. 1442–1452. ISSN: 2168-2267.

[21] Andreas Møgelmose, Antonio Prioletti, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund. "Two-Stage Part-Based Pedestrian Detection". In: *15th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sept. 2012, pp. 73–77.

[22] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Trajectory Analysis and Prediction for improved Pedestrian Safety: Integrated Framework and Evaluations". In: *Intelligent Vehicles Symposium (IV)*. In press. IEEE, June 2015.

[23] Brendan Morris and Mohan M. Trivedi. "Learning and classification of trajectories in dynamic scenes: A general framework for live video analysis". In: *Advanced Video and Signal Based Surveillance, 2008. AVSS'08. IEEE Fifth International Conference on*. IEEE. 2008, pp. 154–161.

[24] Antonio Prioletti, Andreas Møgelmose, Paolo Grisleri, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund. "Part-Based Pedestrian Detection and Feature-Based Tracking for Driver Assistance: Real-Time, Robust Algorithms, and Evaluation." In: *IEEE Transactions on Intelligent Transportation Systems* 14.3 (Sept. 2013), pp. 1346–1359.

[25] Nicolas Schneider and Dariu M. Gavrila. "Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study". In: *Pattern Recognition*. Ed. by Joachim Weickert, Matthias Hein, and Bernt Schiele. Vol. 8142. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 174–183. ISBN: 978-3-642-40601-0.

[26] Junli Tao and Reinhard Klette. "Integrated Pedestrian and Direction Classification Using a Random Decision Forest". In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. Dec. 2013.

[27] Ashish Tawari, Andreas Møgelmose, Sujitha Martin, Thomas B. Moeslund, and Mohan M. Trivedi. "Attention Estimation by Simultaneous Analysis of Viewer and View". In: *17th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2014, pp. 1381–1387.

[28] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* 1 (2001), pp. 511–518. ISSN: 1063-6919.

[29] Paul Viola, Michael J. Jones, and Daniel Snow. "Detecting Pedestrians Using Patterns of Motion and Appearance". In: *International Journal of Computer Vision* 63 (2 2005), pp. 153–161. ISSN: 0920-5691. URL: http://dx.doi.org/10.1007/s11263-005-6644-8.

[30] J. Walker, A. Gupta, and M. Hebert. "Patch to the Future: Unsupervised Visual Prediction". In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. June 2014, pp. 3302–3309.

[31] Dan Xie, S. Todorovic, and Song-Chun Zhu. “Inferring "Dark Matter" and "Dark Energy" from Videos”. In: *Computer Vision (ICCV), 2013 IEEE International Conference on*. Dec. 2013, pp. 2224–2231.

Chapter 4

Person Re-Identification

The final part of this PhD is not about intelligent cars, but related to looking at people and to some of same methods as the pedestrian applications discussed in the previous chapter. This chapter is about person re-identification: recognizing people that a camera-system has observed before. Of particular focus here is trying to introduce multiple modalities other than the well-known visible light RGB modality. As has been the case for the previous chapters, much of the text here is an edited version of excerpts from the included papers [14, 12, 13].

1 Introduction

The task of person re-identification is about recognizing people that have been captured earlier by a camera in a surveillance network. The network may consist of one or more cameras, and can be placed in traditional surveillance contexts or more narrowly scoped areas, such as keeping track of a single queue of people. The objective is simple: When a person enters the field of view of a camera in the system, it must be determined whether or not this person has been seen before.

It is useful in many places where it is desirable to obtain knowledge of the flow of people: airports, transit centers, shopping malls, amusement parks, etc. It can either be knowledge of a single person's movement, or movement patterns in general by combining the patterns of many people. This task can in some cases be solved by a system, which is able to view the entire scene, as in [17, 15]. However, in indoor scenes it is often not feasible to place one camera with a full overview. This is where re-identification enters play. It allows the system designer to place sensors at certain bottlenecks and identify people when they pass these.

Re-identification has the specific distinction from e.g. biometric access

control systems that it must be able to enroll new people on-the-fly and without their specific collaboration. On the other hand, the recognition performance does not necessarily have to be as strong as in access control systems, since re-identification systems are more concerned with the general trend of movement as opposed to the movement of each individual.

The crucial difference between re-identification and tracking is that for re-identification there is expected to be a significant spatial or temporal difference between observations, making it impossible to rely on simple motion dynamics as tracking often does. Instead, soft biometrics are used to decide if a subject has been seen before.

Specifically, a number of challenges and characteristics set re-identification apart from traditional tracking and hard-biometric recognition:

- The set of re-identifiable persons must be updated on the fly; there can be no enrollment phase that requires direct participation from the subjects.
- There is no – or only weak – constraints on the pose of subjects, so the system must be robust to pose changes.
- Persons must be re-identifiable at distances where sensor resolution is generally not sufficient for traditional face recognition.
- The database containing the subjects has a transient nature since subjects are generally not relevant if they have not been re-identified after a certain time span – then they have probably left the area.

Some applications of re-identification do not require all recorded persons to be re-identified. An example is the commercial system from Blip Systems [5], which does person flow tracking in airports based on radio signatures from mobile phones. It has a re-identification rate of around 10%, which is sufficient for a representative flow map.

Traditionally, re-identification has been performed using RGB-video from surveillance networks. However, because the re-identification scenario can be harder than traditional recognition due to the worse data quality, it is an obvious idea to use more sensor modalities. With the advent of the Microsoft Kinect and similar structured light-based sensors (ASUS Xtion and the PrimeSense Sensor), RGB-D sensors have become much more accessible and affordable, and using them in larger surveillance applications does not seem impossible. Similarly, thermal cameras have also become more accessible. Much of the work in this part of the PhD has been centered around trying to introduce these modalities to re-identification.

2 State of the Art

Person re-identification as described above has been an active research area for about a decade and truly gained speed in the latter half of the 2000s. This section describes examples of re-identification systems, but a comprehensive survey on person re-identification can be found in [1].

As mentioned earlier, enrollment in such a system has to happen on the fly. Two types of enrollment strategies exist:

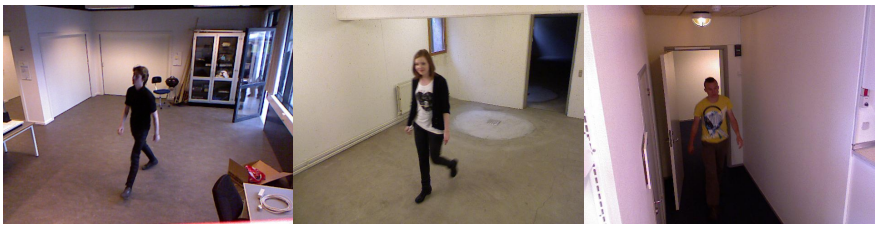
- Single-shot: Enroll and re-identify using a single image.
- Multiple-shot: Enroll and possibly re-identity using a sequence of images.

A single-shot approach is used by Zheng et. al. [20], who split the subject into six stipes, which are each described using color and texture histograms. Their method is feature independent and they propose a new formulation of re-identification as a distance learning problem, where they optimize the probability of a true pair to have a smaller distance than a wrong match pair. Another single shot algorithm is put forth by Zhao et. al. [19]. They compute dense color histograms and SIFT descriptors in patches across the subject bounding box and match these. Before the match, a salience map is made by comparing the subject to a reference database of people (unrelated to the gallery-set in the system). The matching scores are then weighted with the salience scores to ensure that the most descriptive parts of the subject are used. This method seems to be close to the process humans use.

In [2], Bak et. al. proposes a multi-shot approach, where the model structure is learned individually per subject. The best features are selected from a large number of potential features in various color and structural representations of the subject. The learning is performed using an entropy criterion, so each part of the subject is modeled using the optimal descriptor. An earlier multi-shot method was presented by Demirkus et. al. [6], who use both full-body and soft facial biometrics that are directly understandable for people, such as gender, hair color, and clothing color.

Faranzena et. al. [8] describes an approach that can be used in both single-shot and multi-shot frameworks. They describe the overall chromatic contents of the subject, as well as divide them into body parts to gain robustness towards pose changes. Another single- and multi-shot method is that of Hirzer et. al. [10] which first finds and ranks likely matches via a region covariance descriptor, then trains a boosted classifier on these to obtain better performance on that specific subject, a philosophy which is similar to that of [2].

Moving away from the traditional visible light modality, Jüngling and Arens [11], present a full single-shot re-identification pipeline based on in-



(a) (b) (c)

Fig. 4.1: Example images from our own (a) Novi, (b) Basement, and (c) Hallway sequences.

frared images. It detects candidates, then tracks and re-identifies them using SIFT-features. In the depth modality, Barbosa et. al. [4] re-identifies by comparing various physical body measurements (anthropometrics) obtained from the depth image. Velardo and Dugelay [18] uses manually measured anthropometrics to prune the set of candidates for face recognition.

2.1 Datasets

Several public datasets exist, though mostly sets captured with traditional visible light sensors. The two most well known sets are VIPeR [9] and ETHZ [16, 7]. VIPeR contains 632 pedestrian image pairs with an image size of 128x48 pixels. The images are not part of sequences and as such work only with single shot re-identification schemes. They are cropped closely around persons and are thus not relevant for detection either. But for the specific task of re-identifying in RGB, VIPeR is large and generally the best candidate. ETHZ contains images taken from videos, so it can work for multi-shot approaches. It contains 146 persons in sequences of varying durations.

More recent datasets are iLIDS-MA and iLIDS-AA [3]. iLIDS-MA contains 40 manually annotated persons over 2 cameras. There are 46 frames per person per camera, so it works fine for multi-shot evaluations. iLIDS-AA contains 100 persons, but they are annotated automatically (using a HOG-detector), so the bounding boxes are more noisy. The source of these two sets is the Imagery Library for Intelligent Detection Systems (i-LIDS), so the source videos may be obtained.

In other modalities, not many datasets exist. For depth, the RGB-D Person Re-identification Dataset [4] is one option. It contains 79 people in 4 different scenarios: Walking slowly with outstretched arms, two instances of walking from a frontal viewpoint, and walking from a rear viewpoint.

Due to the lack of good multimodel datasets, another dataset was captured during this PhD with a surveillance-like camera setup. Three sequences were captured: Novi, Basement, and Hallway. They all contain sequences of

	Number of persons	Number of frames	Contains image sequences	Available modalities
VIPeR	632	1264	No	RGB
ETHZ	146	8580	Yes	RGB
iLIDS-MA	40	3680	Yes	RGB
iLIDS-AA	100	N/A	Yes	RGB
RGB-D PRD	79	1580	Yes, short	RGB, depth
Novi	22	7800	Yes	RGB, depth
Basement	35	7231	Yes	RGB, depth, thermal
Hallway	10	4492	Yes	RGB, depth, thermal

Table 4.1: Statistics on re-identification datasets

persons walking diagonally towards and past the sensor twice. Novi contains 22 persons over 7800 frames (passes have varying lengths). Basement contains 35 persons over 7231 frames, and Hallway contains 10 persons over 4492 frames. Stats about the public as well as our own datasets can be seen in table 4.1. The sequences were captured with Microsoft Kinect for Xbox. Example pictures from each sequence can be seen in fig. 4.1.

3 Contributions

During this PhD, 3 works have been done with regard to person re-identification. [14] and [12] are linked and both building on the same RGB-D based system. It is a system which relies primarily on color cues for re-identification, but uses the depth modality for easy segmentation of the subjects. It has on-the-fly enrollment and is a full pipeline system from picture input to re-identification. The system was first presented in [14] and later expanded with a more thorough analysis and testing in [12].

In [13], we attempted to extract more features from the depth modality and also added thermal images to the mix. While it seemed intuitively correct that more modalities would result in better re-identification performance, it turned out that the extra information was of very limited use.

Bibliography

[1] "A survey of approaches and trends in person re-identification". In: ().

[2] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. "Learning to Match Appearances by Correlations in a Covariance Metric Space". In: *ECCV (3)*. Vol. 7574. LNCS. Springer, 2012, pp. 806–820. ISBN: 978-3-642-33711-6.

[3] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. "Boosted human re-identification using Riemannian manifolds". In: *Image Vision Comput.* 30.6-7 (2012), pp. 443–452.

[4] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. "Re-identification with RGB-D Sensors." In: *ECCV Workshops (1)*. Vol. 7583. LNCS. Springer, 2012, pp. 433–442. ISBN: 978-3-642-33862-5.

[5] Blip Systems. *Blip Track Airport*. http://www.bliptrack.com/airport/area-of-operations/. 2012.

[6] M. Demirkus, K. Garg, and S. Guler. "Automated person categorization for video surveillance using soft biometrics". In: *Biometric Technology for Human Identification VII*. 2010.

[7] A. Ess, B. Leibe, and L. Van Gool. "Depth and Appearance for Mobile Scene Analysis". In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. Oct. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4409092.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. "Person re-identification by symmetry-driven accumulation of local features". In: *CVPR*. 2010.

[9] *Evaluating Appearance Models for Recognition, Reacquisition, and Tracking*. 2007.

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. "Person Re-identification by Descriptive and Discriminative Classification". In: *SCIA*. Vol. 6688. Lecture Notes in Computer Science. Springer, 2011, pp. 91–102. ISBN: 978-3-642-21226-0.

[11] Kai Jüngling and Michael Arens. "Local Feature Based Person Reidentification in Infrared Image Sequences." In: *AVSS*. IEEE Computer Society, 2010, pp. 448–455. ISBN: 978-0-7695-4264-5.

[12] Andreas Møgelmose, Chris Bahnsen, and Thomas B. Moeslund. "Comparison of Multi-shot Models for Short-term Re-identification of People using RGB-D Sensors". In: *11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP2015)*. Mar. 2015.

[13] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund, and Sergio Escalera. "Tri-modal Person Re-identification with RGB, Depth and Thermal Features". In: *9th Workshop on Perception Beyond the Visible Spectrum, CVPR Workshops*. IEEE, 2013, pp. 301–307.

[14] Andreas Møgelmose, Thomas B. Moeslund, and Kamal Nasrollahi. "Multimodal Person Re-Identification using RGB-D Sensors and a Transient Identification Database". In: *International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2013, pp. 1–4.

[15] Brian E. Moore, Saad Ali, Ramin Mehran, and Mubarak Shah. "Visual crowd surveillance through a hydrodynamics lens". In: *Commun. ACM* 54.12 (2011), pp. 64–73.

[16] W. R. Schwartz and L. S. Davis. "Learning Discriminative Appearance-Based Models Using Partial Least Squares". In: *SIBGRAPI*. IEEE Computer Society, Jan. 26, 2010, pp. 322–329. ISBN: 978-0-7695-3813-6.

[17] Berkan Solmaz, Brian E. Moore, and Mubarak Shah. "Identifying Behaviors in Crowd Scenes Using Stability Analysis for Dynamical Systems". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.10 (2012), pp. 2064–2070.

[18] C. Velardo and J. Dugelay. "Improving Identification by Pruning: A Case Study on Face Recognition and Body Soft Biometric". In: *WIAMIS*. IEEE, 2012, pp. 1–4. ISBN: 978-1-4673-0791-8.

[19] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. "Unsupervised Salience Learning for Person Re-identification". In: CVPR. 2013.

[20] W. Zheng, S. Gong, and T. Xiang. "Person re-identification by probabilistic relative distance comparison". In: *CVPR*. IEEE, 2011, pp. 649–656.

Chapter 5

Conclusion

Running from 2012-2015, this PhD has covered topics in intelligent vehicles and looking at people. Specifically, 3 themes were looked into:

- Traffic sign detection
- Pedestrian detection and analysis
- Person re-identification

In the area of traffic sign detection, a comprehensive survey was made to chart the landscape of previous work. This survey pointed to a number of shortcomings in the state-of-the-art, especially in the area of detecting U.S. traffic signs. To this end, the world's largest collection of U.S. traffic signs pictures was collected and made public for other researchers to use. We also investigated ways to push the state-of-the-art detection performance for U.S. signs and brought the detection performance up to par with that of European signs. The research has shown that regional differences in traffic signs are more severe than was to be expected, as the stellar performance in detecting European signs had somewhat stifled the research interest in the topic. Going forward there are still very relevant open issues, particularly in detecting speed limit signs and rare signs, and also in determining whether detected traffic signs belong to the current road.

For pedestrian detection, a hybrid Haar/HOG approach was suggested and tested with success early in the PhD period. The method combines the speed of the Viola-Jones detection algorithm with the accuracy, and while it has since been surpassed by the state-of-the-art in this very active reseach area, it performed well at the time of publication. Furthermore we have looked into analysing driver behavior combined with pedestrian detection in order to determine which pedestrian a driver might not have seen. A proof-of-concept for such a system was developed. Finally, pedestrian tracking has

been employed in an attempt to predict their future behavior. The tracking, done with a particle filter, worked well, but underlined the need for combining tracking with more subtle motion cues to accurately predict pedestrian behavior.

In person re-identification we have made a push for using multi-modal sensor by adding depth and thermal information to the standard RGB-modality. Multiple iterations of a re-identification system have been developed and tested on a dataset we captured as part of the work. Current results indicate that even in our rather controlled laboratory environment, the additional modalities did not enhance performance, at least the way the system is currently designed. Future work might investigate more advanced combinations of the modalities than the late-fusion approach applied here, and other feature types would be worth investigating as well.

While the PhD-work has been spilt into three somewhat different topics, many of the methods - especially detection algorithms - have been applicable across all areas, and interesting results have been achieved in each of the three topics. The results open up for a wealth of related work, and the traffic sign investigations have already been carried on by others in our research group and applied to traffic signal detection, a challenge which is still lacking behind traffic sign and certainly pedestrian detection.

Part II

Traffic Sign Detection

Paper A

Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey

Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund

The paper has been published in the
IEEE Transactions on Intelligent Transportation Systems Vol. 13.4,
pp. 1484–1497, 2012.

The layout has been revised.

Abstract

In this paper, we provide a survey of the traffic sign detection literature, detailing detection systems for Traffic Sign Recognition (TSR) for driver assistance. We separately describe the contributions of recent works to the various stages inherent in traffic sign detection: segmentation, feature extraction, and final sign detection. While TSR is a well-established research area, we highlight open research issues in the literature, including a dearth of use of publicly-available image databases, and the over-representation of European traffic signs. Further, we discuss future directions for TSR research, including integration of context and localization. We also introduce a new public database containing US traffic signs.

1 Introduction

In this paper, we provide a survey of traffic sign detection for driver assistance. State-of-the-art research utilizes sophisticated methods in computer vision for traffic sign detection and it has been an active area of research over the past decade. On-road applications of vision have included lane detection, driver distraction detection, and occupant pose inference. As described in [63, 62, 61], it is crucial to not only consider the car's surrounding and external environment when designing an assist system, but also to consider the intenal environment and take the driver into account. Fusing other types of information with the sign detector, as described in [40], can make the overall system even better.

When the system is considered a distributed system where the driver is an integral part, it allows for the driver to contribute with what he is good at (e.g. seeing speed limit signs, as we shall see later), while the TSR part can present information from other signs. In addition other surround sensors can also have an influence on what is presented.

In recent years, speed limit detection systems have been included in top of the line models from various manufacturers, but a more general sign detection solution and an integration into other vehicle systems has not yet materialized. Current state-of-the-art TSR systems neither utilize information about the driver, nor input from the driver, to enhance performance. Extensive studies in Human-Machine Interactivity are necessary to present the TSR information in a careful way, to inform the driver without causing distraction or confusion. The literature features just two surveys on TSR: [19] is a good introduction, but not very comprehensive. [18] is a few years old, so any improvements in the field from the past 5 years are not presented. A very good comparison of various segmentation methods is offered in [24], but given that it only covers segmentation, it is not a comprehensive overview of detection

methods. Likewise, [29] provides a good comparison of Hough transform derivatives. In this paper our emphasis is on framing the TSR problem in the context of human-centered driver assistance systems. We provide a comparative discussion of papers published mostly within the last 5 years and to provide an overview of the recent work in the area of sign detection, a subset of the TSR problem.

We provide a critical review of traffic sign detection, and offer suggestions for future research areas in this challenging problem domain. The next section establishes the driver assistance context and covers TSR systems in general. Section 3 provides a problem description, and a gentle introduction to traffic sign detection. Section 4 deals with segmentation for traffic sign detection. Section 5 details models and feature extraction. Section 6 deals with the detection itself. In the final section, the authors provide analysis and insight on future research directions in the field.

2 Human-Centered TSR for Driver Assistance: Issues and Considerations

Traffic sign recognition research needs to take into account the visual system of the driver. This can include factors such as visual saliency of signs, driver focus of attention, and cognitive load. According to [55] (see table A.1 for a summary of the main results), not all signs are equal in their ability to capture the attention of the driver. For example, a driver may fixate his gaze on a sign, but neither notice the sign, nor remember its informational content. While drivers invariably fixate on speed limit signs and recall their information, they are less likely to notice game crossing and pedestrian signs. This can endanger pedestrians, as it may not leave enough reaction time to stop.

The implications of use of TSR in human-in-the-loop system are clear; instead of focusing on detection and recognizing all signs of some class perfectly, which would be the objetive for an autonomous car, the task is now to detect and highlight signs that the driver has not seen. This gives way to various models of TSR, which take into account the driver's focus of attention, and interactivity issues. Driver attention tracking is covered in [17] and [42]. Fig. A.1 presents examples on how TSR can be used for driver assistance. Fig. A.1a shows how a system should act in an autonomous car. It simply recognizes all signs present. In fig. A.1b there is a driver in the loop, and while the system may see all the signs, it should avoid presenting them in order to avoid driver confusion. Instead, it simply highlights the sign type that is easy to overlook, like the pedestrian crossing warnings in the research. Fig. A.1c shows how a driver is distracted by a passing car. This causes him

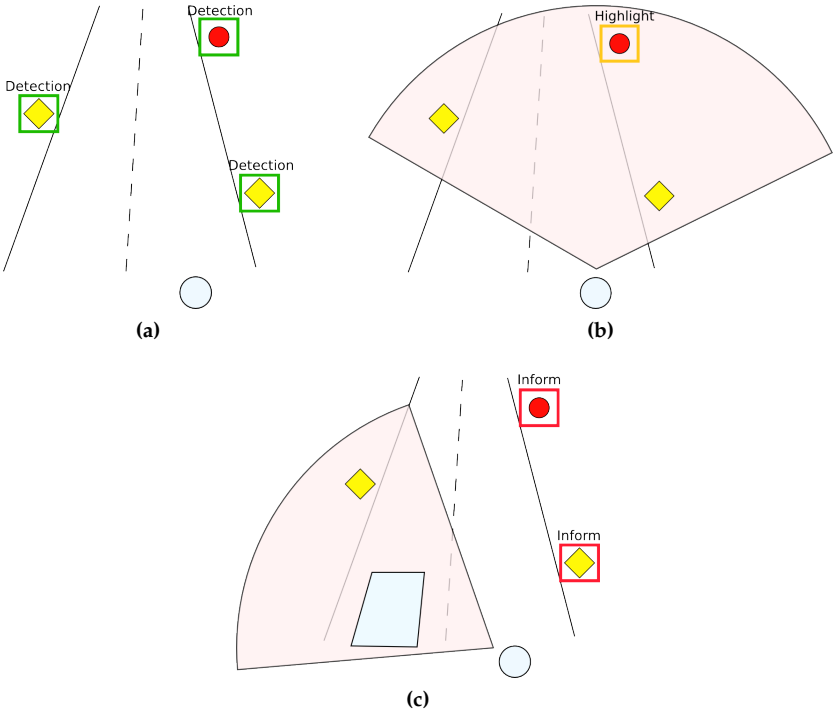


Fig. A.1: Different detection scenarios. The circle is the ego-car and 3 signs are distributed along the road. The area highlighted in red illustrates the driver's area of attention. (a) is the standard scenario used for e.g. autonomous cars. Here, all signs must be detected and processed. (b) and (c) depicts a system which tracks the driver's attention. In (b), the driver is attentive and spots all signs. Therefore the system just highlights the one sign that is known to be difficult for people to notice. In (c), the driver is distracted by a passing car and thus misses two signs. In this case, the system should inform the driver about the two missed signs.

to miss two signs. His car has a TSR system for driver assistance, which informs him of the signs as he returns his attention to the road ahead of him. This could, for example, be done using a heads-up display as suggested in [15].

Even though this paper is mostly concerned with using TSR for driver assistance, TSR has various well defined applications, summarized nicely by [13]:

1. Highway maintenance: Check the presence and condition of signs along major roads.

2. Sign inventory: Similar to the above task, create an inventory of signs in city environments.

Table A.1: Significant results from [55] regarding attention to various sign types.

	Fixated		Not fixated	
Target	Recalled	Not recalled	Recalled	Not recalled
Speed limit 80 km sign	100	0	0	0
Game Crossing sign	60	0	7	33
Pedestrian crossing ahead	8	54	0	38
Pedestrian crossing sign	0	21	0	79

3. Driver assistance systems: Assist the driver by informing of current restrictions, limits, and warnings.

4. Intelligent autonomous vehicles: Any autonomous car that is to drive on public roads must have a means of obtaining the current traffic regulations. This can be done through TSR.

This paper uses the term TSR to refer to the entire chain from detection of signs to their classification, and potentially presentation to the driver. Generally, TSR is split into two stages: Detection and classification (see fig. A.2). Detection is concerned with locating signs in input images, while classification is about determining what type of sign the system is looking at. The two tasks can often be treated as completely separate, but in some cases the classifier relies on the detector to supply information, such as the sign shape or sign size. In a full system, the two stages are depending on each other and it does not make sense to have a classifier without a detection stage. Later, we divide the detection stage into three sub-stages, but these should not be confused with the two main stages of a full TSR-system: Detection and classification.

Apart from shape and color, another aspect may be used in TSR: Temporal information. Most TSR systems are designed with a video feed from a vehicle in mind, so signs can be tracked over time. The simplest way of using tracking is to accept sign candidates as signs only if they have shown up on

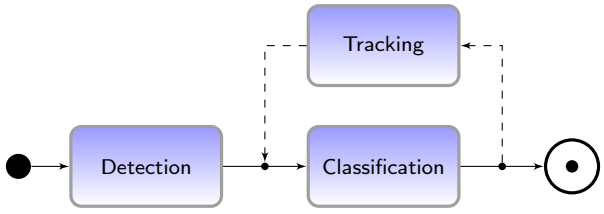


Fig. A.2: The basic flow in most TSR systems.

a number of consecutive frames. Sign candidates that only show up once are usually a result of noise. Employing a predictive method, such as a Kalman filter, allows for the system to predict where a sign candidate should show up in the next frame, and if its position is too far away from this prediction, the sign candidate is discarded. A predictive tracking system has the additional benefit of handling occlusions, hence preventing signs that were occluded from being classified as new signs. This is very important in a driver assistance system where signs should only be presented once, and in a consistent way. Imagine a scenario where a sign is detected in a few frames and the occluded for a short while, before being detected again. For an autonomous car it is likely not a problem to be presented with the same information twice: If the first sign prompted the speed to be set at 55 mph, there is no problem in the system being told once again that the speed limit is 55 mph. In a driver assistance system, the system must not present more information than absolutely necessary at any given moment, so the driver is not overwhelmed with information, for instance, forcing the driver to pay attention to a sign he has already seen should be avoided.

Many TSR systems are tailored to a specific sign type. Due to the vast differences in sign design from region to region (see the following section), and the differences in sign design based on their purpose, many systems narrow their scope down to a specific sign type in a specific country.

There is a wide span in speeds of the systems. For use in driver assistance and autonomous vehicles, real-time performance is necessary. This does not necessarily mean a speed of 30 Hz, but the signs must be read quickly enough to still be relevant to act on. Depending on the exact application, a few Hz is required.

Instead of treating the entire TSR-process in what could easily become a cursory manner, we have opted to look thoroughly on the detection stage. The line between detection and classification is a bit blurry, since some detectors provide more information to the classifier than others. It is normal for the detector to inform the classifier of the general category of signs, since that is often defined by either the overall sign shape or its color, something that the detector itself may use to to localize the sign.

Even though this paper is targeted towards the problem of detecting traffic signs, one must not forget that without a subsequent classification stage, the systems are useless. So even though we encourage a decoupling of the two tasks, this does not mean that the classification is a solved problem. It is a crucial part of a full system.

Fig. A.3: Examples of European signs. These are Danish, but many countries use similar signs. (a) Speed limit. Sign C55. (b) End speed limit. Sign C56. (c) Start of freeway. Sign E55. (d) Right turn. Sign A41.

3 Traffic signs

Traffic signs are markers placed along roads to inform drivers about either road conditions and restrictions or which direction to go. They communicate a wealth of information, but are designed to do so efficiently and at a glance. This also means that they are often designed to stand out from their surroundings, making the detection task fairly well defined.

The designs of traffic signs are standardized through laws, but differ across the world. In Europe many signs are standardized via the Vienna Convention on Road Signs And Signals [64]. There, shapes are used to categorize different types of signs: Circular signs are prohibitions including speed limits, triangular signs are warnings and rectangular signs are used for recommendations or sub-signs in conjunction with one of the other shapes. In addition to these, octagonal signs are used to signal a full stop, downwards pointing triangles yield, and countries have other different types, e.g. to inform about city limits. Examples of these signs can be seen in fig. A.3.

In the US, traffic signs are regulated by the Manual on Uniform Traffic Control Devices (MUTCD) [58]. It defines which signs exist and how they should be used. It is accompanied by the Standard Highway Signs and Markings (SHSM) book, which describes the exact designs and measurements of signs. At the time of writing, the most recent MUTCD was from 2009, while the SHSM book had not been updated since 2004, and thus it described the MUTCD from 2003. The MUTCD contains a few hundred different signs, divided into 13 categories.

To complicate matters further, each US state can decide whether it wishes to follow the MUTCD. A state has three options:

1. Adopt the MUTCD fully as is.
2. Adopt the MUTCD but add a State Supplement.
3. Adopt a State MUTCD that is “in substantial conformance with” the national MUTCD.

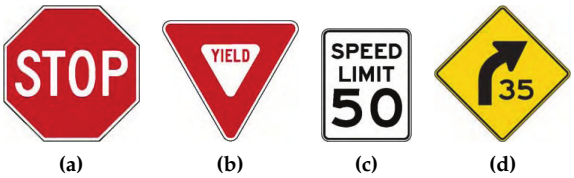


Fig. A.4: Examples of signs from the US national MUTCD. (a) Stop. Sign R1-1. (b) Yield. Sign R1-2. (c) Speed limit. Sign R2-1. (d) Turn warning with speed recommendation. Sign W1-2a.. Image source: [58]

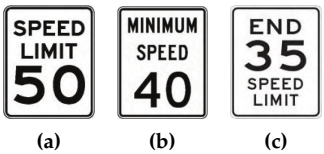


Fig. A.5: Examples of similar signs from the MUTCD. (a) Speed limit. Sign R2-1. (b) Minimum speed. Sign R2-4. (c) End speed limit. Sign R3 (CA), exists only in the California MUTCD. Image source: [58]

In the US 19 states have adopted the national MUTCD without modifications, 23 have adopted the national MUTCD with a state supplement and 10 have opted to create a State MUTCD (the count includes the District of Columbia and Puerto Rico). Examples of US signs can be seen in fig. A.4.

New Zealand uses a sign standard with warning signs that are yellow diamonds, as in the US, but regulatory signs that are round with a red border, like the ones from the Vienna Convention countries. Japan uses signs that are generally in compliance with the Vienna Convention, as are Chinese regulatory signs. Chinese warning signs are triangular with a black/yellow color scheme. Central and South American countries do not participate in any international standard, but often use signs somewhat like the American standard.

While signs are well defined through laws and designed to be easy to spot, there are still plenty of challenges for TSR systems. They include:

- Signs being similar within or across categories (see fig. A.5).
- Signs may have faded or be dirty so they are no longer their specified color.
- The sign post may be bent, so the sign is no longer orthogonal to the road.

- Lighting conditions may make color detection unreliable.
- Low contrast may make shape detection hard.
- In cluttered urban environments, other objects may look very similar to signs.
- Varying weather conditions.

3.1 Assessing performance of sign detectors

When comparing sign detectors, some comparison metrics must be set up. The straight forward and most important measure is the true positive rate. However, even if all signs are detected, the system is not necessarily perfect. The number of false positives must also be taken into account. If the amount of false positives is too high, the classifier will have to handle a lot more data than it should, degrading the overall system speed. For cases when a system must work in real-time in a car, obviously the detection must be fast. In general, the faster the detection runs, the more time is left over for the classification stage. Adjusting these goals is a trade-off. Often, the target will be to create a system that is just fast enough for a given application, while keeping the receiver operating characteristic acceptable. Another interesting performance characteristic is what sign types a given system works for.

Even with the parameters in mind, and a clear idea of the performance metrics, comparing the performance of different systems is not a straightforward task. Unlike other computer vision areas, until recently no standardized training and test data set existed, so no two systems were tested with the same data. The image quality varies from high resolution still images (as in [65, 21, 59]) to low resolution frames from in-car video cameras (such as [67, 53, 30]). That, combined with the facts that signs vary wildly between countries, and many papers limit their scope to specific sign types, makes for a quite uneven playing field. For a discussion of the performance of the papers presented in this survey, see section 4

3.2 Public sign databases

A few publicly available traffic sign datasets exist:

- German Traffic Sign Recognition Benchmark (GTSRB) [57, 56]
- KUL Belgium Traffic Signs Dataset (KUL Dataset) [60]
- Swedish Traffic Signs Dataset (STS Dataset) [34]
- RUG Traffic Sign Image Database (RUG Dataset) [26]
- Stereopolis Database [7]

Table A.2: Information on the publicly available sign databases.

	GTSRB	**STS Dataset**	**KUL Dataset**	**RUG Dataset**	**Stereopolis**	**LISA Dataset**
Number of classes:	43	7	100+	3	10	49
Number of annotations:	50000+	3488	13444	0	251	7855
Number of images:	50000+	20000	9006	48	847	6610
Annotated images:	All images	4000 images	All images	0	All images	All images
Sign sizes:	15x15 to 250x250 px	3x5 to 263x248 px	100x100 to 1628x1236 px	N/A	25x25 to 204x159 px	6x6 to 167x168 px
Image sizes:	15x15 to 250x250 px	1280x960 px	1628x1236 px	360x270 px	1920x1080 px	640x480 to 1024x522 px
Includes videos:	No	No	Yes, 4 tracks	No	No	Yes, for all annotations
Country of origin:	Germany	Sweden	Belgium	The Netherlands	France	United States
Extra info:	Images come in tracks with 30 different images of the same physical sign.	Signs marked visible/blurred/occluded and whether they belong to the current road or a side road.	Includes traffic sign annotations, camera calibrations and poses.	Does not include any annotations, only raw pictures.		Images from various camera types.

Information on these databases can be found in table A.2. Most of the databases have emerged within the last two years (except for the very small RUG Dataset), and are not yet widely used. One of the most widespread databases is the GTSRB, which has been presented in [57], created for the competition "The German Traffic Sign Recognition Benchmark". The competition was held at the International Joint Conference on Neural Networks (IJCNN) 2011. It is a large data set containing German signs, thus very suitable for training and testing systems aimed at signs adhering to the Vienna Convention. A sample image from the GTSRB database can be found in fig. A.6a. The GTSRB is primarily geared towards classification, rather than detection, since each image contains exactly one sign without much background. For detection, images of complete scenes is necessary. Also, many detection systems rely on a tracking scheme to make detection more robust and without video of the tracks (in GTSRB parlance a "track" is a set of images of the same physical sign), this will not work properly. Since the data set is created for the classification task, this is not so much a problem of that database, as it is a testament to its target. In conjunction with the competition, five interesting papers [49, 11, 69, 54, 9] were released. They all focus on classification rather than detection.

Two other datasets should be highlighted: The STS Dataset and the KUL Dataset. They are both very large, though not as large as the GTSRB, and they contain full images. This means that they can both be used for detection purposes. The STS Dataset does not have all images annotated, but it does include all frames from the videos used to obtain the data. This means that tracking systems can be used on this dataset, but it can only be verified with ground truth every 5 frames. An example from the STS Dataset can be seen in fig. A.6b. The KUL Dataset also includes 4 recorded sequences which can be used for tracking experiments. KUL also includes a set of sign-free images which can be used as negative training images and it has pose-information for the cameras for each image.

From the research it was evident that there was a lack of databases with US traffic signs, so in conjunction with this paper we have assembled one. Its details are also listed in table A.2. One novel feature of this dataset is that it includes video tracks of all the annotated signs. Many systems already use various tracking schemes to minimize the number of false positives, and it is quite likely that in the future, detectors using temporal data even more will emerge. Therefore, the LISA dataset includes video as well as stand alone frames. Not all frames have been extracted for annotation, but all annotated frames can be traced back to the source video so the annotations can also be used to verify systems using tracking.

(a) **(b)**

Fig. A.6: Example sign images from (a) the GTSRB and (b) the STS Dataset with the sign bounding boxes superimposed.

4 Sign detection

The approaches in this stage have traditionally been divided into two kinds:

- Color based methods.
- Shape based methods.

Color based methods take advantage of the fact that traffic signs are designed to be easily distinguished from their surroundings, often colored in highly visible contrasting colors. These colors are extracted from the input image and used as a base for the detection. Just like signs have specific colors, they also have very well defined shapes that can be searched for. Shape based methods ignore the color in favor of the characteristic shape of signs.

Each method has its pros and cons. Color of signs, while well defined in theory, varies a lot with available lighting, as well as with age and condition of the sign. On the other hand, searching for specific colors in an image is fairly straight forward. Sign shapes are invariant to lighting and age, but parts of the sign can be occluded, making the detection harder, or the sign may be located at a background of a similar color, ruining the edge detection that most shape detectors rely on.

The division of systems in this way can be problematic. Almost all color based approaches take shape into account after having looked at colors. Others use shape detection as their main method, but integrate some color aspects as well. Instead, the detection can be split into two steps as proposed by [24]: Segmentation and detection. In this paper we go one step further and split the detection step into a feature extraction step and the actual detection, which acts on the features that are extracted. Many shape-only based

Fig. A.7: The general flow followed by typical sign detection algorithms.

methods have no segmentation step. The flow is outlined in fig. A.7.

An overview of all surveyed papers and their methods is listed in table A.3 (these large tables can be found towards the end of the chapter). It contains each of the systems and lists which segmentation method, feature type, and detection method that is used. The author group numbers are used to mark the papers that are part of an ongoing effort from the same group of authors. They do not constitute a ranking in any way. In tables A.4 and A.5, some of their more detailed properties are listed. The systems are split into two tables. Table A.4 displays those which do not use any tracking. Table A.5 contain those which do use tracking, something we find crucial when using TSR in a driver assistance context, as mentioned earlier. Apart from this division, the two tables are structured in the same way: *Sign type in paper* describes which sign types the authors of the paper have attempted to find, while emphsign type possible are the types of signs the method could be extended to include, usually a very broad group. *Real-time* is about how fast the system runs, if that information is available. Any system with a frame rate faster than 5 fps is considered to have real-time potential. *Rotation invariance* tells whether the used technique is robust to rotation of signs. *Model vs. training* describes if the detection system relies on a theoretical model of signs (such as a pre-defined shape), if it uses a learned type of classifier, or if it uses a combination of the two. *Test image type* is the image resolution the system is designed to work with. Low-res images are usually video frames, while high-res are still images.

The detection performance of the surveyed papers are presented in table A.6. As mentioned earlier, very few papers use common databases to test their performance and the papers detect various types and numbers of signs. Thus, the numbers should not be directly compared, but nevertheless they give an idea of performance. Not all papers report all the measures reported in the table (detection rate, false positives per frame, etc.), so some fields in the table could not be filled. In other cases these exact measures were not given, but could be calculated from other given numbers. Where figures are available, the best detection rate the system obtained is reported along with the corresponding measure of false positives. The detection rate is per frame, meaning that 100% detection is only achieved if a sign is found in every frame it is present. It is not sufficient to just detect the sign in a few frames. This is

the way results are presented in most papers, so this is the measure chosen here, even if a real-world system would work fine if each sign is just detected once. Papers which only report the per-sign detection rate as opposed to the per-frame detection rate are marked with a triangle in the right-most column of the table.

Different papers report the false positives in different ways, so a few different measures - which are not directly comparable - are presented in the table:

FPPF False postives per frame: $FPPF = \frac{FP}{f}$ where FP is the number of false positives and f is the number of frames analyzed.

FPR False positive rate: $FPR = \frac{FP}{N}$ where N is the number of negatives in the test set. This measure is rarely used in detection, since the number of negatives does not always make much sense (how many negatives exist in a full frame?).

PPV Positive predictive value: $PPV = \frac{TP}{TP+FP}$ where TP is the number of true positives.

FPTP False/true positive ratio: $FPTP = \frac{FP}{TP}$

WPA Wrong pixels per area: $WPA = \frac{WP}{AP}$ where WP is the number of wrongly classified pixels and AP is the total number of pixels classified.

When papers present results for different sign types, the mean detection performance is also presented in the table. In many cases that will give a better view of the true performance of the approach.

Five papers stick out, claiming a 100% detection rate. The first [23] is only tested on synthetic data. It is possible that the synthetic data does not fully encapsulate real world variations, so the performance of that approach is not guaranteed to be as good in real-world scenarios. At first glance [51] achieves a 100% detection rate, but that is only the case for one of their sign types. The mean performance is a more accurate (and still promising) gauge of the actual performance. The same is the case for [34]. [36] detects all signs in the test set, but at the cost of a large number of false positives per frame. [28] only presents the per-sign detection rate, so that figure cannot be compared to the other systems.

Generally, systems achieve detection rates well into the 90% range, some at very low false detection rates. From the table no "best system" can be chosen, since the test sets are very different, both in size and content. A system that can detect several different sign types at a low detection rate may in some applications be considered better than a system that can only detect

one specific sign type, but do that very well. A few papers that should be highlighted are [59, 6, 27, 45]. They have all been tested on large datasets and report detection rates above 90% with a decent low number of false positives.

Now that the basics about sign detection are in place, the following sections go in depth with how recent papers perform each step.

5 Segmentation

The purpose of the segmentation step is to achieve a rough idea about where signs might be, and thus narrow down the search space for the next steps. Not all authors make use of this step. Since the segmentation is traditionally done based on colors, authors who believe this should not be part of a sign detection often do not have any segmentation step, but go directly to the detection.

Of the papers that do use segmentation, all except [27, 31] use colors to some extent. Normally, segmentation is done with colors and subsequently a shape detection is run in a later stage. In [27], the usual order is reversed, so they use radial symmetry voting (see section 7) for segmentation and a color based approach for the detection. [31] also run radial symmetry voting as preprocessing, but follow it up with a cascaded classifier using Haar wavelets (see again section 7).

Generally, color based segmentation relies on a thresholding of the input image in some color space. Since many believe that the RGB color space is very fragile with regards to changes in lighting, these methods are spearheaded by the HSI-space (or its close sibling, the HSV-space). HSI/HSV is used by [32, 43, 50, 68, 10, 48]. The HSI-space models the human vision better than RGB and allows some variation in the lighting, most notably in the intensity of light. Some papers, like the ones in the series starting with [65] and followed by [38, 23, 33], augment the HSI thresholding with a way to find white signs. Hue and saturation are not reliable for detecting white, since it can be at any hue, so they use an achromatic decomposition of the image proposed by [35].

Some authors are not satisfied with the performance of HSI, since it does not model the change in color temperature in different weather, but only helps in changing light intensity. [21, 20] instead threshold in the LCH color space, which is obtained using the CIECAM97-model. This allows them to take variations in color temperature into account. The RGB space is used by [59, 47], but they use an adaptive threshold in an attempt to combat instabilities caused by lighting variations.

Of special interest in this color space discussion is the excellent paper [24], which has shown that HSI-based segmentation offers no significant ben-

(a) Before thresholding **(b)** After thresholding

Fig. A.8: An example of thresholding, looking for red hues.

efit over normalized RGB, but that methods which use color segmentation generally perform much better than shape-only methods. They do, however have trouble with white signs. For a long time, it has simply been assumed that the RGB color space was a bad choice for segmentation, but through rigorous testing, they show that there is nothing to gain from switching to the HSI color space instead of a normalized RGB space. As the authors write: "Why use a nonlinear and complex transformation if a simple normalization is good enough?".

A color based model not relying on thresholding was put forward in [14], which use a cascaded classifier trained with AdaBoost, similar to the one proposed by [66], but on Local Rank Pattern features instead of Haar wavelets. Also, [51] use a color-based search method that, while closely related to, is not directly thresholding-based. Here, the image is discretized into colors that may exist on signs. The discretization process is less destructive than thresholding in that it does not directly discard pixels, instead it maps them into the closest sign-relevant color. In a more recent contribution [53], they replace the color discretization method with a Quad-tree interest region finding algorithm, which finds interesting areas using an iterative search method for colored signs. In the same realm lies [29], which uses a learned probabilistic color preprocessing.

In [30], a unique approach is proposed: Using a biologically inspired attention system. It produces a heat map denoting areas where signs are likely to be found. An example can be seen in figure A.9. A somewhat similar system was put forth by [67], who uses a saliency measure to find possible areas of interests.

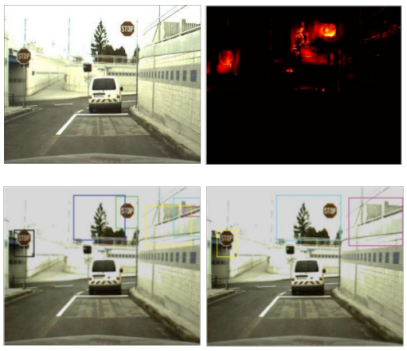


Fig. A.9: The biologically inspired detection stage from [30]. Image source: [30]

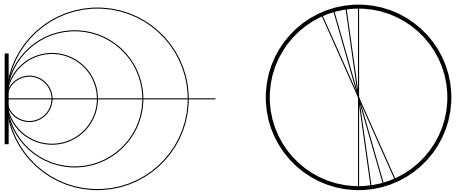
6 Features and modeling

While various features are available from the vision literature, the choice of feature set is often closely coupled with the detection method, though some feature sets can be used with a selection of different detection methods. The most popular feature is edges - sometimes edges obtained directly from the raw picture, sometimes edges from pre-segmented images. Edges are practically always found using a Canny edge detection or some method very similar, and they are used as the only feature in [51, 53, 36, 3, 5, 44, 39, 22, 25, 59, 35, 32, 41, 8, 50, 10, 48, 14, 29]. [47] combine the edges with Haar-like features and [52, 28] look only at certain color filtered edges.

Even though edges comprise the most popular feature choice, there are other options. Histogram of Oriented Gradients (HOG) is one. It was first used to detect people in images, but has been used by [1, 46, 45, 21, 67] to detect signs. HOG is based on creating histograms of gradient orientations on patches of the image and comparing them to known histograms for the sought after objects. HOG is also used in [12], but they augment the HOG feature vectors with color information to make them even more robust.

A number of papers [2, 31, 47, 6] use Haar wavelet-like features, [2] only on certain colors, and [6] in the form of so-called dissociated dipoles with wider structure options than traditional Haar wavelets.

More esoteric choices are Distance to Bounding box (DtB), FFT of shape signatures, tangent functions, simple image patches, and combinations of various simple features. DtB, as used in [38, 33], are a measure of distances from the contour of a sign-candidate to its bounding box. Similarly, the FFT of shape signatures used in [23] is based on the distance from the shape center to its contour at different angles. Tangent functions, used in [68], calculate the angles of the tangents at various points around the contour. Simple image patches (though in the YCbCr color space) are championed by [43] and a



(a) Possible circles for a gradient

(b) Intersecting vote lines

Fig. A.10: The basic principle behind the radial symmetry detector. Image inspired by [5].

combination of simple features, such as corner positions and color is used in [30].

7 Detection

The detection stage is where the signs are actually found. This is in many ways the most critical step, and often also the most complicated. The selection of detection method is a bit more constrained than the previous two stages, since the method must work with the features from the previous stage. The decision is therefore often made the other way around: A desired detection method is chosen, and the feature extraction stage is designed to deliver what is necessary to perform the detection. As we know from the previous section, the most popular feature is edges, and this reflects on the most popular choice in detection method. Using Hough transforms to process the edges is one option, as done by [41, 22, 25, 50]. In [41], a proprietary and undisclosed algorithm is used for detection of rectangles in addition to the Hough transform used for circles. That said, Hough transforms are computationally expensive and not suited for systems with real-time requirements. Because of that, the most popular methods are derivatives of the radial symmetry detector first proposed in [37] and first put to use for sign detection in [4]. The algorithm votes for the most likely sign centers in an image based on symmetric edges and is itself inspired by the Hough transform. The basic principle can be seen in fig. A.10. In a circle, all edge gradients intersect at the center. The algorithm finds gradients with a magnitude above a certain threshold. In the direction pointed out by the gradient, it casts a vote in a separate vote image. It looks for circles of a specific radius and thus votes only in the distance from the edge that is equivalent to the radius. The places with most votes are most likely to be the center of circles.

This algorithm was later extended to regular polygons by [36] and a faster implementation for sign detection use was proposed by [3]. It is also used in some form by [5, 31, 39, 28, 27, 44]. An example of votes from a system

Fig. A.11: Votes from a radial symmetry system superimposed to the original image. The brightest spot coincides with the center of the sign. This image is from a system developed in conjunction with this paper and is a radial symmetry voting algorithm extended to work for rectangles

which is extended to work for rectangular signs can be seen on fig. A.11. An alternate edge-based voting system is proposed by [8].

The HOG features can be used with an SVM, as in [67, 12], or be compared by calculating a similarity coefficient as in [21]. Another option with regard to HOG is to use a cascaded classifier trained with some type of boosting. This is done in [46, 45]. Cascaded classifiers are traditionally used with Haar wavelets, and sign detection is no exception, as used in [2, 31, 47, 6].

Finally, also neural networks and genetic algorithms are represented in [43] and [35], respectively.

The detection stage reflects the philosophical difference that was also seen in the feature extraction stage: Either reliance on a simple, theoretical model of sign shapes is preferred - at this stage it is nearly always shapes that are searched for - or reliance on training data and then a more abstract detection method. Since it is extremely hard to compare systems tested across different data sets, it is not clear which methods perform the best, so that is clearly an area with a need for further studies. Both ways can be fast enough for real-time performance, and most of them could also work with signs of any shape. There are outliers using different methods, but no compelling argument that they should perform significantly better.

8 Discussion and future directions

In the previous sections, different methods and philosophies for each stage are presented. This section discusses the current state of the art and outlines ideas for future directions of research.

At the moment, the problem in TSR is the lack of use of standardized sign image databases. This makes comparisons between contributions very hard. In order to obtain meaningful advances in the field, the development of such databases is crucial. Until now, research teams have only implemented a method they believe has potential, or perhaps tested a few solutions. Without a way to compare performance with other systems, it is not clear which approaches work the best, so every new team starts back at square one, implementing what *they* think might work best. Two efforts to remedy this situation deserve to be mentioned: The sign databases presented earlier and the segmentation evaluation in [24]. As mentioned earlier (section 3.2), a few public sign databases have recently emerged, but have not yet been widely used. In [24], the authors compare various segmentation methods on the same data set containing a total of 552 signs in 313 images. They also propose a way to evaluate the performance of segmentation methods. That paper provides a very good starting point for determining which segmentation method to use.

These two efforts notwithstanding, public databases covering signs from non-Vienna Convention regions are necessary. Databases which include video tracks of signs would also be very beneficial to the development of TSR systems, since many detectors employ a tracking system for signs. This is, to some extent, included in the KUL Dataset. In relation to the work on this present survey, we have assembled such a database for US traffic signs, one that includes full video tracks of signs. It is our hope that the GTSRB database will also be extended to include video and full frames and that more US databases will be created.

The absence of usage of public database may not explain in entirety why very few comparative studies of methods exist. Another reason is that TSR systems are long, complex chains of various methods, where it is not always possible to swap individual modules. When it is not feasible to swap, say, the detection method for something else, it is naturally hard to determine whether other solutions may be better. This is solved, if more papers divide their work more clearly into stages, ideally as fine grained as the ones used in this survey, plus a similar set of stages for classification. This is done with success in [24], as they test different segmentation methods while keeping the feature extraction, detection, and classification stages fixed.

Fig. A.12: Example of sign relevancy challenges in a crop from our own collected data set. The signs have been manually highlighted, and while both signs would likely be detected, only the one to the right is relevant to the driver. The sign to the left belongs to another road, the one the black and white cars come from.

Another problem is the need for work on TSR in regions not adhering to the Vienna Convention. The bulk of the existing work comes out of Europe, Australia, and Japan. Japan and Australia are not parts of the Vienna Convention, but they use similar signs, for example to convey speed limits. Of the surveyed papers here, only two are concerned with US traffic signs [31, 41], and even they only look at speed limit signs.

When looking at sign detection from a driver-in-the-loop perspective, it is also unfortunate that the bulk of reseach now focuses on speed limit signs. A wealth of papers cite driver assistance as their main application, but carries on focusing on speed limit signs. Detection of speed limits is highly relevant for an autonomous vehicle, but as it turns out, humans are already very good at seeing speed limit signs themselves [55]. As such, recognition of signs other than speed limit is actually more interesting.

The final problem we wish to highlight in this section is the relation of signs to the surroundings. TSR has seen significant work, as is evident from this paper, but little work has been done on ensuring that the detected signs are relevant for the ego-car (with the notable exception of [22]). In many situations, it can occur that a detected sign is not connected to the road the car is on. An example from our own collected data can be seen in fig. A.12. In this case, two stop signs can be seen, but only the rightmost one pertains to the current road. Similar situations occur often on freeways, where some signs may only be relevant for exit lanes. Related to this problem is that when the driver changes to a different road, most often the restrictions from earlier detected signs no longer apply. This should be detected and relayed to the

system. It is very likely that research in other areas, such as lane detection can be of benefit here. Another idea with regard to the surroundings would be to link knowledge of weather and current lighting conditions to enhance the robustness of the detector, similar to what is done for detection of people in [16]. It is also possible that vehicle dynamics can be taken into account and used in the tracking of detected signs.

9 Concluding remarks

This paper provides an overview of the state of sign detection. Instead of treating the entire TSR flow, focus has been solely on the detection of signs. During recent years, a large effort has gone into TSR, mainly from Europe, Japan, and Australia and the developments have been described.

The detection process has been split into segmentation, feature extraction, and detection. Many segmentation approaches exist, mostly based on evaluating colors in various color spaces. For features there are also a wealth of options. The choice is made in conjunction with the choice of detection method. By far the most popular features are edges and gradients, but other options such as HOG and Haar wavelets have been investigated. The detection stage is dominated by the Hough transform and its derivatives, but for HOG and Haar wavelet features, SVMs, neural networks, and cascaded classifiers have also been used.

Arguably, the biggest issue with sign detection as it is currently is the lack of use of public image databases to train and test systems. Currently, every new approach presented uses a new dataset for testing, making comparisons between papers hard. This gives the TSR effort a somewhat scattered look. Recently, a few databases have been made available, but they are still not widely used, and cover only Vienna Convention compliant signs. We have contributed with a new database, the LISA Dataset, which contains US traffic signs.

This issue leads to the main unanswered question in sign detection: Is a model based shape detector superior to a learned approach, or vice versa? Systems using both approaches exist, but are hard to compare, since they all use different data sets.

Many contributions cite driver assistance systems as their main motivation for creating the system, but so far only little effort has gone into the area of combining TSR systems with other aspects of driver assistance and notably, none of the studies include knowledge about the driver's behavoir in order to tailor the performance of the TSR system to the driver.

Other open issues include lack of research in finding non-European style signs and detected signs are hard to relate to their surroundings.

Acknowledgment

The authors would like to thank our colleagues in the LISA-CVRR lab, especially Mr. Sayanan Sivaraman, Mr. Minh Van Ly, Ms. Sujitha Martin, and Mr. Eshed Ohn-Bar for their comments.

Table A.3: Overview of detection methods in 41 recent papers. Papers with the same background color are papers written by the same group. White background indicate stand-alone papers.

Paper	Year	Author group	Segmentation method	Features	Detection method
[65]	2005	1	HSI thresholding with addition for white signs ([35])	Boundary distance transform	Correlation with model distance transforms
[38]	2007	1	HSI thresholding with addition for white signs ([35])	DtB (distance to bounding box)	Linear SVM
[23]	2008	1	HSI thresholding with addition for white signs ([35])	FFT of shape signatures	Euclidian nearest neighbor
[33]	2010	1	HSI thresholding with addition for white signs ([35])	DtB (distance to bounding box)	Linear SVM
[2]	2005	2	None	Haar wavelet features computed on specific color channels	Cascaded classifier
[31]	2008	2	Extended radial symmetry voting	Haar wavelet features	Cascaded classifier
[21]	2006	3	LCH thresholding (obtained with CIECAM97)	HOG	Comparison with template vectors
[20]	2008	3	LCH thresholding (obtained with CIECAM97)	None	None
[52]	2007	4	None	Color filtered edges	Extended radial symmetry voting
[51]	2010	4	HSV discretization	Edges	Extended radial symmetry voting
[53]	2011	4	Quad-tree color selection	Edges	Extended radial symmetry voting
[36]	2004	5	None	Edges	Extended radial symmetry voting
[3]	2006	5	None	Edges	Extended radial symmetry voting
[5]	2008	5	None	Edges	Radial symmetry voting
[44]	2008	6	None	Edges	Votes for symmetric areas to be used as ROI with another shape-detector
[39]	2011	6	None	Edges	Two-tier radial symmetry voting
[22]	2011	7	None	Edges of closed contours with certain aspect ratios	Hough shape detection

Continued on next page...

Continued from previous page...

Paper	Year	Author group	Segmentation method	Features	Detection method
[25]	2011	7	None	Edges of closed contours with certain aspect ratios	Hough shape detection
[59]	2009	8	Adaptive RGB threshold	Edges	Fuzzy templates (a Hough derivative)
[47]	2010	8	Adaptive RGB threshold	Edges and Haar-like features	Fuzzy templates, cascaded classifier, and SVM
[46]	2008	9	None	HistFeat (HOG derived)	Cascaded classifier
[45]	2011	9	None	Various HOG-features	5 stage cascaded classifier trained with LogitBoost
[35]	2002	None	HSI thresholding with edge detection and removal of achromatic colors	Edges	Genetic algorithm looking for circles
[1]	2007	None	None	Edge orientation histograms	Comparison with template vectors
[32]	2007	None	HSI thresholding	Edges	Hough shape detection
[41]	2007	None	None	Edges	Hough transform for circular signs, proprietary (not described) for rectangular
[43]	2008	None	HSI thresholding	30x30 px YcbCr patches	Neural network
[6]	2009	None	None	Dissociated dipoles	Cascaded classifier
[8]	2009	None	None	Edges	Vertex and Bisector transform (VBT)
[28]	2009	None	Radial symmetry voting combined with SIFT features	Edge colors	Contracting Curve Density
[50]	2009	None	HSV thresholding	Edges	Hough shape detection
[67]	2009	None	Saliency detection with color and edges	HOG	SVM
[68]	2009	None	Hue thresholding on chromatic colors only	Tangent function of simplified contours	Distance from model tangent function
[10]	2010	None	HSI thresholding	Edges	Circle center voting
[12]	2010	None	None	HOG augmented with color information	SVM
[30]	2010	None	Biologically inspired attention model	Color, corner positions, height, excentricity	Color, corner positions, height, excentricity
[48]	2010	None	HSI thresholding	Edges	Radial symmetry voting

Continued on next page...

Continued from previous page...

Paper	Year	Author group	Segmentation method	Features	Detection method
[14]	2011	None	Nested cascade classifier with Local Rank Pattern features (based on 7 RGB based colors)	Edges	RANSAC circle fit
[27]	2011	None	Radial symmetry voting	Colors in modified RGB-space	Distance to learned colors
[29]	2011	None	Probabilistic color preprocessing	Edges	Hough derivative shape detector
[34]	2011	None	None	Fourier descriptors	Correlation based matching

Table A.4: Overview of detailed properties of the 27 papers which do not use tracking

Paper	Year	Author group	Sign type in paper	Sign type possible	Real-time	Rotation inv.	Model vs. training	Test image type
[65]	2005	1	5 regular polygons, various colors	Colored	N/A	Yes	Both	High-res
[23]	2008	1	Circular, triangular, square and semiellipses	Colored	N/A	Yes	Both	Low-res
[33]	2010	1	Circular red	Circular colored	No	No	Both	Low-res
[21]	2006	3	Circular red, circular blue and triangular red	Colored	No	Yes	Both	High-res
[20]	2008	3	Circular red and blue	Colored	N/A	Yes	Training	High-res
[52]	2007	4	Circular, triangular and square, various colors	Regular, colored polygons	Yes	No	Both	N/A
[36]	2004	5	Regular polygons	Regular polygons	Yes	Yes	Model	Low-res
[3]	2006	5	Regular polygons	Regular polygons	Yes	No	Model	Low-res
[5]	2008	5	Circular red	Circular	Yes	Yes	Model	Low-res
[59]	2009	8	Circular red, circular blue, diamond white	Colored	No	No	Both	High-res
[46]	2008	9	Circular, triangular and octagonal red	Any sign	Yes	No	Training	Low-res
[45]	2011	9	Circular red	Any sign	Yes	No	Training	Low-res
[35]	2002	None	Circular, red	Colored	N/A	Yes	N/A	N/A
[1]	2007	None	Circular, triangular, diamond, octagonal	Any sign	Yes	No	Training	Low-res
[32]	2007	None	Circular and triangular, red	Colored	N/A	Yes	N/A	N/A
[43]	2008	None	Circular red, triangular red, and octagonal red	Colored	No	No	Training	N/A
[6]	2009	None	Circular and triangular, red	Any sign	No	No	Training	N/A
[8]	2009	None	Triangular red and blue	Triangular	Yes	Yes	Model	Low-res
[50]	2009	None	Circular, triangular, square, various colors	Colored	No	Yes	Model	Low-res
[67]	2009	None	Circular, red and square blue	Colored	N/A	Yes	Training	Low-res

Continued on next page...

Continued from previous page...

Paper	Year	Author group	Sign type in paper	Sign type possible	Real-time	Rotation inv.	Model vs. training	Test image type
[68]	2009	None	Circular, triangular, square, various colors	Colored	N/A	No	Both	Low-res
[10]	2010	None	Circular red	Colored	No	No	Model	Low-res
[12]	2010	None	Circular red, circular blue and triangular red	Any sign	N/A	No	Training	High-res
[30]	2010	None	Triangular and octagonal, red	Any sign	Yes	N/A	Model	N/A
[48]	2010	None	Circular red	Colored	N/A	No	Both	Low-res
[29]	2011	None	Circular and triangular, red and blue	Regular, colored polygons	No	No	Both	High-res
[34]	2011	None	7 sign types	Any sign	N/A	Yes	Model	Low-res

Table A.5: Overview of detailed properties of the 14 papers which use tracking

Paper	Year	Author group	Sign type in paper	Sign type possible	Real-time	Rotation inv.	Model vs. training	Test image type
[38]	2007	1	Circular and triangular, red	Colored	No	No	Both	Low-res
[2]	2005	2	Circular red	Any sign	Yes	No	Training	Low-res
[31]	2008	2	Rectangular white	Any sign	Yes	No	Both	Low-res
[51]	2010	4	40 different signs	Any sign	Yes	No	Model	Low-res
[53]	2011	4	Circular red and blue	Regular, colored polygons	Yes	No	Both	Low-res
[44]	2008	6	Circular, triangular and octagonal	Any sign	Yes	No	Model	N/A
[39]	2011	6	Circular	Circular	Yes	Yes	Model	Low-res
[22]	2011	7	100 different signs, circular and triangular	Any sign	Yes	Yes	Model	Low-res
[25]	2011	7	8 sign categories	Any sign	No	Yes	Model	Low-res
[47]	2010	8	N/A	Colored	Yes	N/A	Both	Low-res
[41]	2007	None	Circular red and rectangular white	Circular and rectangular	N/A	N/A	Model	Low-res
[28]	2009	None	Circular red	Circular	N/A	Yes	Both	Low-res
[14]	2011	None	Circular red	Colored	No	Yes	Training	Low-res
[27]	2011	None	Circular red and blue	Circular colored	Yes	Yes	Both	Low-res

Table A.6: Overview of the performance of the papers included in this survey. For those papers where the numbers are available, the best and mean detection rate is presented, along with the corresponding false positive measure. Note that the systems have all been tested in different ways, so a direct comparison is not feasible. See section 4 for further details.

Paper	Year	Group	Evaluation data format	Pos/neg in evaluation data	Best detection rate	False positives for best detection	Mean detection rate	Mean false positives
[65]	2005	1	No statistical results given					
[38]	2007	1	5176 images from 5 videos containing 104 signs	N/A	N/A	N/A	N/A	N/A
[23]	2008	1	2000 synthetic images	N/A	100%	WPA: 0.74%	89.08%	WPA: 13.17%
[33]	2010	1	No statistical results given					
[2]	2005	2	Images from videos	1700 pos/40000 neg	98.6%	FPR: 0.03%	-	-
[31]	2008	2	16828 images from videos	80 positives	98.75%	FPPF: 0.062	-	-
[21]	2006	3	Images from videos	98 positives	95%	N/A	-	-
[20]	2008	3	128 images	142 positives	94%	PPV: 23%	89.67%	PPV: 26%
[52]	2007	4	No results given for the detection stage only					
[51]	2010	4	Images from videos containing 210 signs	N/A	100%	N/A	92.9%	N/A
[53]	2011	4	No statistical results given (graphs are available in the paper)					
[36]	2004	5	45 images	49 positives	100%	FPPF: 0.67	96.67%	FPPF: 0.56
[3]	2006	5	47 images from 1 video	47 positives	93.62%	FPPF: 2.26	-	-
[5]	2008	5	Images from videos	N/A	93%	FPPF: 0.5	-	-
[44]	2008	6	No statistical results given (graphs are available in the paper)					
[39]	2011	6	Images from 34 videos containing more than 100 signs	N/A	87.12%	FPR: 0.14%	-	-
[22]	2011	7	30000 images from 1 video	340 positives	97.74%	FPPF: 0.0024	96.45%	FPPF: 0.0014
[25]	2011	7	Images from videos containing 500 signs	N/A	99.96%	N/A	99.52%	N/A

Continued on next page...

Continued from previous page...

Paper	Year	Group	Evaluation data format	Pos/neg in evaluation data	Best detection rate	False positives for best detection	Mean detection rate	Mean false positives	
[59]	2009	8	7356 images containing 269 signs	2459 positives	95.7%	FPPF: 2.5	-	-	
[47]	2010	8	No statistical results given (see [59] instead)						
[46]	2008	9	No statistical results given (graphs are available in the paper)						
[45]	2011	9	Images from videos	21500 pos/40000 neg	98.68%	FPR: 10^{-8}%	-	-	
[35]	2002	None	No statistical results given						
[1]	2007	None	Video tracks, 10-200 frames in length	105 positives	81.9%	N/A	-	-	
[32]	2007	None	No results given for the detection stage only						
[41]	2007	None	Images from videos containing 281 signs	N/A	88.97%	FPPF: 0	-	-	
[43]	2008	None	164 images	164 positives	92.45%	N/A	-	-	
[6]	2009	None	4755 images from 4 videos	N/A	97%	FPPF: 0.056	92%	FPPF: 0.048	
[8]	2009	None	48 images	40 positives	82.5%	FPPF: 0.042	-	-	
[28]	2009	None	Images from 1 30 min. video containing 94 signs	N/A	100%	N/A	-	-	△
[50]	2009	None	Images from videos containing 20 signs	N/A	N/A	N/A	-	-	
[67]	2009	None	More than 500 images from videos	N/A	99.16%	FPR: 5.56%	98.3%	FPR: 4.72%	
[68]	2009	None	1000 images	N/A	95%	FPTP: 0%	91.8%	FPTP: 0.9%	
[10]	2010	None	Images from videos	397 pos/697 neg	89.42%	FPR: 0.05%	-	-	
[12]	2010	None	3000 images	N/A	85%	N/A	72.47%	N/A	
[30]	2010	None	820 images from 2 videos	117 positives	89.8%	PPV: 98.3%	-	-	
[48]	2010	None	85 images from video	95 positives	76.64%	FPPF: 0.094	-	-	

Continued on next page...

Continued from previous page...

Paper	Year	Group	Evaluation data format	Pos/neg in evaluation data	Best detection rate	False positives for best detection	Mean detection rate	Mean false positives
[14]	2011	None	2967 images from videos	4886 positives	90.1%	PPV: 85.6%	-	-
[27]	2011	None	2134 images from videos	3298 positives	94.03%	FPPF: 3.41	-	-
[29]	2011	None	Comparison of different methods, thus no final result to report.					
[34]	2011	None	STS dataset	641 positives	95.33%	PPV: 100%	77.08%	PPV: 91.85%

Bibliography

[1] B. Alefs, G. Eschemann, H. Ramoser, and C. Beleznai. "Road Sign Detection from Edge Orientation Histograms". In: *Intelligent Vehicles Symposium, 2007 IEEE*. June 2007, pp. 993–998. DOI: 10.1109/IVS.2007.4290246.

[2] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler. "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information". In: *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*. IEEE. 2005, pp. 255–260.

[3] N. Barnes and G. Loy. "Real-time regular polygonal sign detection". In: *Field and Service Robotics*. Springer. 2006, pp. 55–66.

[4] N. Barnes and A. Zelinsky. "Real-time radial symmetry for speed sign detection". In: *Intelligent Vehicles Symposium, 2004 IEEE*. June 2004, pp. 566–571. DOI: 10.1109/IVS.2004.1336446.

[5] N. Barnes, A. Zelinsky, and L.S. Fletcher. "Real-Time Speed Sign Detection Using the Radial Symmetry Detector". In: *Intelligent Transportation Systems, IEEE Transactions on* 9.2 (June 2008), pp. 322–332. ISSN: 1524-9050. DOI: 10.1109/TITS.2008.922935.

[6] X. Baro, S. Escalera, J. Vitria, O. Pujol, and P. Radeva. "Traffic Sign Recognition Using Evolutionary Adaboost Detection and Forest-ECOC Classification". In: *Intelligent Transportation Systems, IEEE Transactions on* 10.1 (Mar. 2009), pp. 113–126. ISSN: 1524-9050. DOI: 10.1109/TITS.2008.2011702.

[7] R. Belaroussi, P. Foucher, J.P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis. "Road sign detection in images: A case study". In: *20th International Conference on Pattern Recognition (ICPR)*. 2010.

[8] R. Belaroussi and J.-P. Tarel. "Angle vertex and bisector geometric model for triangular road sign detection". In: *Applications of Computer Vision (WACV), 2009 Workshop on*. Dec. 2009, pp. 1–7. DOI: 10.1109/WACV.2009.5403030.

[9] F. Boi and L. Gagliardini. "A Support Vector Machines network for traffic sign recognition". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE. 2011, pp. 2210–2216.

[10] Hsin-Han Chiang, Yen-Lin Chen, Wen-Qing Wang, and Tsu-Tian Lee. "Road speed sign recognition using edge-voting principle and learning vector quantization network". In: *Computer Symposium (ICS), 2010 International*. Dec. 2010, pp. 246–251. DOI: 10.1109/COMPSYM.2010.5685511.

[11] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber. "A committee of neural networks for traffic sign classification". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE. 2011, pp. 1918–1921.

[12] I.M. Creusen, R.G.J. Wijnhoven, E. Herbschleb, and P.H.N. de With. "Color exploitation in hog-based traffic sign detection". In: *Image Processing (ICIP), 2010 17th IEEE International Conference on*. Sept. 2010, pp. 2669–2672. DOI: 10.1109/ICIP.2010.5651637.

[13] A. De la Escalera, J.M. Armingol, and M. Mata. "Traffic sign recognition and analysis for intelligent vehicles". In: *Image and vision computing* 21.3 (2003), pp. 247–258.

[14] D. Deguchi, M. Shirasuna, K. Doman, I. Ide, and H. Murase. "Intelligent traffic sign detector: Adaptive learning based on online gathering of training samples". In: *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE. 2011, pp. 72–77.

[15] A. Doshi, S.Y. Cheng, and Mohan M. Trivedi. "A novel active heads-up display for driver assistance". In: *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 39.1 (2009), pp. 85–93.

[16] A. Doshi and Mohan M. Trivedi. "Satellite imagery based adaptive background models and shadow suppression". In: *Signal, Image and Video Processing* 1.2 (2007), pp. 119–132.

[17] Anup Doshi and Mohan M. Trivedi. "Attention estimation by simultaneous observation of viewer and view". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE. 2010, pp. 21–27.

[18] H. Fleyeh and M. Dougherty. "Road and traffic sign detection and recognition". In: *10th EWGT Meeting and 16th Mini-EURO Conference*. 2005, pp. 644–653.

[19] Meng-Yin Fu and Yuan-Shui Huang. "A survey of traffic sign recognition". In: *Wavelet Analysis and Pattern Recognition (ICWAPR), 2010 International Conference on*. July 2010, pp. 119–124. DOI: 10.1109/ICWAPR.2010.5576425.

[20] Xiaohong Gao, Kunbin Hong, Peter Passmore, Lubov Podladchikova, and Dmitry Shaposhnikov. "Colour vision model-based approach for segmentation of traffic signs". In: *J. Image Video Process.* 2008 (Jan. 2008), 6:1–6:7. ISSN: 1687-5176. DOI: 10.1155/2008.

[21] X.W. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, and N. Shevtsova. "Recognition of traffic signs based on their colour and shape features extracted using human vision models". In: *Journal of Visual Communication and Image Representation* 17.4 (2006), pp. 675–685.

[22] M. A. Garcia-Garrido, M. Ocana, D. F. Llorca, M. A. Sotelo, E. Arroyo, and A. Llamazares. "Robust traffic signs detection by means of vision and V2I communications". In: *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. Oct. 2011, pp. 1003–1008. DOI: 10.1109/ITSC.2011.6082844.

[23] P. Gil Jiménez, S.M. Bascón, H.G. Moreno, S.L. Arroyo, and F.L. Ferreras. "Traffic sign shape classification and localization based on the normalized FFT of the signature of blobs and 2D homographies". In: *Signal Processing* 88.12 (2008), pp. 2943–2955.

[24] H. Gomez-Moreno, S. Maldonado-Bascon, P. Gil-Jimenez, and S. Lafuente-Arroyo. "Goal Evaluation of Segmentation Algorithms for Traffic Sign Recognition". In: *Intelligent Transportation Systems, IEEE Transactions on* 11.4 (Dec. 2010), pp. 917–930. ISSN: 1524-9050. DOI: 10.1109/TITS.2010.2054084.

[25] A. Gonzalez, M.A. Garrido, D.F. Llorca, M. Gavilan, J.P. Fernandez, P.F. Alcantarilla, I. Parra, F. Herranz, L.M. Bergasa, M.A. Sotelo, and P. Revenga de Toro. "Automatic Traffic Signs and Panels Inspection System Using Computer Vision". In: *Intelligent Transportation Systems, IEEE Transactions on* 12.2 (June 2011), pp. 485–499. ISSN: 1524-9050. DOI: 10.1109/TITS.2010.2098029.

[26] C. Grigorescu and N. Petkov. "Distance sets for shape filters and shape recognition". In: *Image Processing, IEEE Transactions on* 12.10 (2003), pp. 1274–1286.

[27] Yanlei Gu, T. Yendo, M.P. Tehrani, T. Fujii, and M. Tanimoto. "Traffic sign detection in dual-focal active camera system". In: *Intelligent Vehicles Symposium (IV), 2011 IEEE*. June 2011, pp. 1054–1059. DOI: 10.1109/IVS.2011.5940513.

[28] B. Hoferlin and K. Zimmermann. "Towards reliable traffic sign recognition". In: *Intelligent Vehicles Symposium, 2009 IEEE*. June 2009, pp. 324–329. DOI: 10.1109/IVS.2009.5164298.

[29] S. Houben. "A single target voting scheme for traffic sign detection". In: *Intelligent Vehicles Symposium (IV), 2011 IEEE*. June 2011, pp. 124–129. DOI: 10.1109/IVS.2011.5940429.

[30] R. Kastner, T. Michalke, T. Burbach, J. Fritsch, and C. Goerick. "Attention-based traffic sign recognition with an array of weak classifiers". In: *Intelligent Vehicles Symposium (IV), 2010 IEEE*. June 2010, pp. 333–339. DOI: 10.1109/IVS.2010.5548143.

[31] C.G. Keller, C. Sprunk, C. Bahlmann, J. Giebel, and G. Baratoff. "Real-time recognition of U.S. speed signs". In: *Intelligent Vehicles Symposium, IEEE*. June 2008, pp. 518–523. DOI: 10.1109/IVS.2008.4621282.

[32] Wen-Jia Kuo and Chien-Chung Lin. "Two-Stage Road Sign Detection and Recognition". In: *Multimedia and Expo, 2007 IEEE International Conference on*. July 2007, pp. 1427–1430. DOI: 10.1109/ICME.2007.4284928.

[33] S. Lafuente-Arroyo, S. Salcedo-Sanz, S. Maldonado-Bascón, J. A. Portilla-Figueras, and R. J. López-Sastre. "A decision support system for the automatic management of keep-clear signs based on support vector machines and geographic information systems". In: *Expert Syst. Appl.* 37 (1 Jan. 2010), pp. 767–773. ISSN: 0957-4174.

[34] F. Larsson and M. Felsberg. "Using fourier descriptors and spatial models for traffic sign recognition". In: *Image Analysis* (2011), pp. 238–249.

[35] H. Liu, D. Liu, and J. Xin. "Real-time recognition of road traffic sign in motion image based on genetic algorithm". In: *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*. Vol. 1. IEEE. 2002, pp. 83–86. DOI: 10.1109/ICMLC.2002.1176714.

[36] G. Loy and N. Barnes. "Fast shape-based road sign detection for a driver assistance system". In: *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*. Vol. 1. IEEE. 2004, pp. 70–75.

[37] G. Loy and A. Zelinsky. "Fast radial symmetry for detecting points of interest". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 25.8 (Aug. 2003), pp. 959–973. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2003.1217601.

[38] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. López-Ferreras. "Road-sign detection and recognition based on support vector machines". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.2 (2007), pp. 264–278.

[39] M. Meuter, C. Nunn, S. M. Gormer, S. Muller-Schneiders, and A. Kummert. "A Decision Fusion and Reasoning Module for a Traffic Sign Recognition System". In: *Intelligent Transportation Systems, IEEE Transactions on* 12.4 (Dec. 2011), pp. 1126–1134. DOI: 10.1109/TITS.2011.2157497.

[40] B. Morris and Mohan M. Trivedi. "Vehicle Iconic Surround Observer: Visualization platform for intelligent driver support applications". In: *Intelligent Vehicles Symposium (IV), 2010 IEEE*. IEEE. 2010, pp. 168–173.

[41] F. Moutarde, A. Bargeton, A. Herbin, and L. Chanussot. "Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system". In: *Intelligent Vehicles Symposium*. IEEE. 2007, pp. 1122–1126.

[42] E. Murphy-Chutorian, A. Doshi, and Mohan M. Trivedi. "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation". In: *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*. IEEE. 2007, pp. 709–714.

[43] Yok-Yen Nguwi and Abbas Kouzani. "Detection and classification of road signs in natural environments". In: *Neural Computing & Applications* 17 (2008), pp. 265–289. ISSN: 0941-0643.

[44] C. Nunn, A. Kummert, and S. Muller-Schneiders. "A two stage detection module for traffic signs". In: *International Conference on Vehicular Electronics and Safety, 2008 (ICVES)*. Sept. 2008, pp. 248–252. DOI: 10.1109/ICVES.2008.4640901.

[45] G. Overett and L. Petersson. "Large scale sign detection using HOG feature variants". In: *Intelligent Vehicles Symposium (IV), 2011 IEEE*. June 2011, pp. 326–331. DOI: 10.1109/IVS.2011.5940549.

[46] N. Pettersson, L. Petersson, and L. Andersson. "The histogram feature - a resource-efficient Weak Classifier". In: *Intelligent Vehicles Symposium, 2008 IEEE*. June 2008, pp. 678–683. DOI: 10.1109/IVS.2008.4621174.

[47] V.A. Prisacariu, R. Timofte, K. Zimmermann, I. Reid, and L. Van Gool. "Integrating Object Detection with 3D Tracking Towards a Better Driver Assistance System". In: *20th International Conference on Pattern Recognition (ICPR)*. Aug. 2010, pp. 3344–3347. DOI: 10.1109/ICPR.2010.816.

[48] Xu Qingsong, Su Juan, and Liu Tiantian. "A detection and recognition method for prohibition traffic signs". In: *Image Analysis and Signal Processing (IASP), 2010 International Conference on*. Apr. 2010, pp. 583–586. DOI: 10.1109/IASP.2010.5476048.

[49] R. Rajesh, K. Rajeev, K. Suchithra, V.P. Lekhesh, V. Gopakumar, and N.K. Ragesh. "Coherence vector of Oriented Gradients for traffic sign recognition using Neural Networks". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. 31 2011-aug. 5 2011, pp. 907–910. DOI: 10.1109/IJCNN.2011.6033318.

[50] FeiXiang Ren, Jinsheng Huang, Ruyi Jiang, and R. Klette. "General traffic sign recognition by feature matching". In: *Image and Vision Computing New Zealand, 2009. IVCNZ '09. 24th International Conference*. Nov. 2009, pp. 409–414. DOI: 10.1109/IVCNZ.2009.5378370.

[51] A. Ruta, Y. Li, and X. Liu. "Real-time traffic sign recognition from video by class-specific discriminative features". In: *Pattern Recognition* 43.1 (2010), pp. 416–430.

[52] A. Ruta, Y. Li, and X. Liu. "Towards real-time traffic sign recognition by class-specific discriminative features". In: *Proc. of the 18th British Machine Vision Conference*. Vol. 1. 2007, pp. 399–408.

[53] A. Ruta, F. Porikli, S. Watanabe, and Y. Li. "In-vehicle camera traffic sign detection and recognition". In: *Machine Vision and Applications* 22 (2 2011), pp. 359–375. ISSN: 0932-8092. DOI: 10.1007/s00138-009-0231-x.

[54] P. Sermanet and Y. LeCun. "Traffic sign recognition with multi-scale Convolutional Networks". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE. 2011, pp. 2809–2813.

[55] David Shinar. *Traffic safety and human behaviour*. Emerald Group Publishing, 2007. ISBN: 978-0-08-045029-2.

[56] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition". In: *Neural Networks* (2012), pp. 323–332. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.02.016. URL: http://www.sciencedirect.com/science/article/pii/S0893608012000457.

[57] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "The German Traffic Sign Recognition Benchmark: A multi-class classification competition". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE. 2011, pp. 1453–1460.

[58] State of California, Department of Transportation. *California Manual on Uniform Traffic Control Devices for Streets and Highways*.

[59] R. Timofte, K. Zimmermann, and L. Van Gool. "Multi-view traffic sign detection, recognition, and 3d localisation". In: *Applications of Computer Vision (WACV), 2009 Workshop on*. Ieee. 2009, pp. 1–8.

[60] Radu Timofte, Karel Zimmermann, and Luc Van Gool. "Multi-view Traffic Sign Detection, Recognition, and 3D Localisation". English. In: *Machine Vision and Applications* (Dec. 2011), pp. 1–15. ISSN: 0932-8092. DOI: 10.1007/s00138-011-0391-3.

[61] Cuong Tran and Mohan M. Trivedi. "Vision for Driver Assistance: Looking at People in a Vehicle". In: *Guide to Visual Analysis of Humans: Looking at People*. Ed. by Thomas B. Moeslund, L. Sigal, V. Krueger, and A. Hilton. 2011. ISBN: 978-0-85729-996-3.

[62] Mohan M. Trivedi and S.Y. Cheng. "Holistic sensing and active displays for intelligent driver support systems". In: *Computer* 40.5 (2007), pp. 60–68.

[63] Mohan M. Trivedi, Tarak Gandhi, and Joel McCall. "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.1 (2007), pp. 108–120.

[64] United Nations Economic Commission for Europe. *Convention on Road Signs And Signals, of 1968*. 2006.

[65] A. Vázquez-Reina, S. Lafuente-Arroyo, P. Siegmann, S. Maldonado-Bascón, and FJ Acevedo-Rodríguez. "Traffic sign shape classification based on correlation techniques". In: *Proceedings of the 5th WSEAS International Conference on Signal Processing, Computational Geometry & Artificial Vision*. World Scientific, Engineering Academy, and Society (WSEAS). 2005, pp. 149–154.

[66] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* 1 (2001), pp. 511–518. ISSN: 1063-6919.

[67] Yuan Xie, Li-Feng Liu, Cui-Hua Li, and Yan-Yun Qu. "Unifying visual saliency with HOG feature learning for traffic sign detection". In: *Intelligent Vehicles Symposium, 2009 IEEE*. June 2009, pp. 24–29. DOI: 10.1109/IVS.2009.5164247.

[68] S. Xu. "Robust traffic sign shape recognition using geometric matching". In: *Intelligent Transport Systems, IET* 3.1 (Mar. 2009), pp. 10–18. ISSN: 1751-956X. DOI: 10.1049/iet-its:20070058.

[69] F. Zaklouta, B. Stanciulescu, and O. Hamdoun. "Traffic sign classification using K-d trees and Random Forests". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. 31 2011-aug. 5 2011, pp. 2151–2155. DOI: 10.1109/IJCNN.2011.6033494.

Paper B

Learning to Detect Traffic Signs: Comparative Evaluation of Synthetic and Real-World Datasets

Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund

The paper has been published in the
Proceedings of the 21st International Conference on Pattern Recognition (ICPR),
pp. 3452–3455, 2012.

The layout has been revised.

Abstract

This study compares the performance of sign detection based on synthetic training data to the performance of detection based on real-world training images. Viola-Jones detectors are created for 4 different traffic signs with both synthetic and real data, and varying numbers of training samples. The detectors are tested and compared. The result is that while others have successfully used synthetic training data in a classification context, it does not seem to be a good solution for detection. Even when the synthetic data covers a large part of the parameter space, it still performs significantly worse than real-world data.

1 Motivation

With the emergence of more advanced sensors embedded in cars, the field of Traffic Sign Recognition (TSR) has seen increasing interest over the last decade. TSR systems can be used in a number of scenarios, ranging from Advanced Driver Assistance Systems (ADAS) - as described in [14] - to fully autonomous cars.

Many sign detection systems (see section 2) rely on large amounts of training data to work. Over the past two years, a few traffic sign datasets has shown up: The GTSRB dataset [13, 12], the Swedish Traffic Signs Dataset [6], and the KUL Belgium Traffic Signs Dataset. A commonality among these datasets is that they contain European Signs conforming to the Vienna Convention. Since signs differ from region to region and in many cases from country to country, an interesting proposition is to use synthetically generated training data, saving a lot of time and effort in gathering the data. Synthetic training data has not yet been widely used in the field of TSR, but is worth researching since very few datasets from outside of Europe exist. A recent survey [9] shows that research on the detection and recognition of traffic signs outside countries conforming to the Vienna Convertion on traffic signs is lacking in general. This paper investigates if using synthetic data for the detection of traffic signs is feasible.

The role and importance of high quality, representative datasets in the development of TSR systems cannot be overemphasized. Collection of such datasets is an expensive task (in time as well as effort). Issues in training, annotations in the real-world, and semi-supervised learning for object recognition is treated in [11]. Since signs have a well-defined appearance, the idea of using synthetic data emerge. The use of synthetic training in sign detection is not yet widespread, prompting this paper. Our paper is focused closely on the generation of synthetic training data for detection purposes. It is also the first of its kind dealing with US signs. In [4, 3], generation of synthetic data specifically for classification is investigated. In [10], some aspects of detecting non-US signs with synthetic data is discussed. The detection task is somewhat harder the classification due to the lack of knowledge about whether a sign is present, where it is, and what size it has.

The following section briefly covers the general workings of TSR systems, followed by a section on how we generate synthetic training data. Towards the end of the paper, the performance of synthetic training data is compared to the performance of real-world training data when used to train a simple AdaBoost cascade with Haar-like features [15].

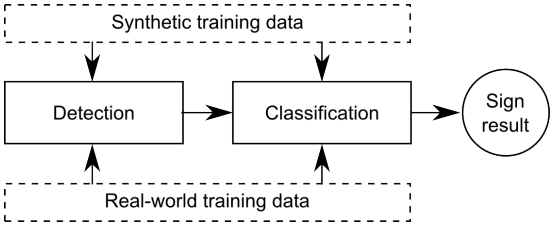


Fig. B.1: Flow for ML-based TSR-systems. The stages can be trained with synthetic or real-world data, and two stages does not have to be trained with the same type.

2 TSR: General approaches

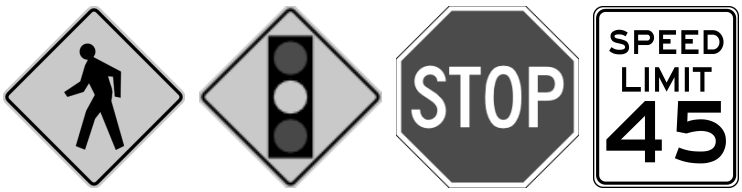
Overviews of TSR can be found in [9, 8, 2]. TSR can be split into two main stages: Sign detection and sign classification, as seen in fig. B.1. Not all detection approaches require training as such, since they are using a theoretical model of the sign, based on e.g. the shape. With that said, many papers present ML based approaches. In [1], an AdaBoost Cascade similar to the one used in this paper was used, albeit on specific color channels. In [5], the image is segmented with a HSI threshold and then classifies the resulting blobs using a linear SVM on DtB features. DtB features are measurements of the distance between the edge of the blob and its rectangular bounding box.

3 Synthetic training data for detection

The question this paper tries to answer is: Can we substitute real-world training data with synthetic in ML based sign detection systems? The idea is to generate synthetic training images from a drawn template. Template examples can be seen in fig. B.2.

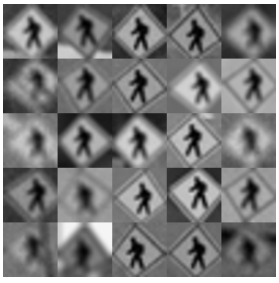
The goal is to emulate how signs of the given type might look on pictures from the real world. In order to do this, several transformations are made randomly to the template:

Hue variations emulates faded signs and color casts due to lighting of the natural

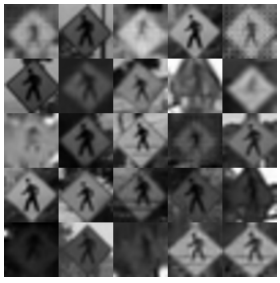


(a) Pedestrian crossing **(b)** Signal ahead **(c)** Stop **(d)** Speed limit

Fig. B.2: Examples of typical US sign templates.



(a) Synthetic training images. Template in fig. B.2a



(b) Real-world training images.

(c) Synthetic training images. Template in fig. B.2c

(d) Real-world training images.

Fig. B.3: Samples from the training image sets.

Table B.1: Results of the comparative evaluations of detectors

Training type	Training images (positive/negative)	Stages	Signs to find	TP	FP	FN
Stop						
Real-world	1218/2500	20	103	76 (73.8%)	11	27
Real-world	1686/3000	20	103	75 (72.8%)	8	28
Synthetic	1218/2500	17	103	18 (17.5%)	2	85
Synthetic	5000/10000	19	103	26 (25.2%)	5	77
Synthetic	1218/2500	10	103	60 (58.3%)	1500	43
Pedestrian crossing						
Real-world	364/800	20	40	29 (72.5%)	10	11
Real-world	1044/2000	20	40	30 (75%)	2	10
Synthetic	364/800	14	40	11 (27.5%)	28	29
Speed limit 35						
Real-world	253/500	20	21	15 (71.4%)	1	6
Synthetic	253/500	7	21	5 (23.8%)	32	16
Synthetic	2000/4000	7	21	6 (28.6%)	6	15
Signal ahead						
Real-world	597/1500	20	56	42 (75%)	10	14
Real-world	859/2000	20	56	38 (67.9%)	4	18
Synthetic	597/1500	13	56	14 (25%)	117	49
Synthetic	2000/4000	13	56	16 (28.6%)	53	48

scene. Done by adding to/subtracting from the hue-parameter in the HSV color space.

Lighting variations emulates shadows and variations in exposure. Done by adding to/subtracting from the value-parameter in the HSV color space.

Rotations around the x-, y-, and z-axis with the origin in the center of the template. Emulates signs captured from different perspectives.

Backgrounds taken from a real image are added to the template. This emulates the various backgrounds a sign might have in real life.

Gaussian blur is added to emulate an unfocused camera. It should be noted that Gaussian blur does not really emulate the bokeh produced by an unfocused lens, but emulating bokeh properly is not a straightforward task, and it would likely not give any notable detection benefit.

Gaussian noise to emulate sensor noise.

Occlusions are added in the form of tree branches growing in front of some signs.

Each transformation should be applied with a random parameter within some realistic boundaries. Samples of training images can be seen in fig. B.3.

To evaluate whether the synthetic datasets cover the same variance in appearance as the real-world data, we compare the distributions in intensity- and blur-values among training sets. In fig. B.4a a plot of the mean of the intensities in the training images is shown. Each point in the plot is a single image. Data for the detectors of

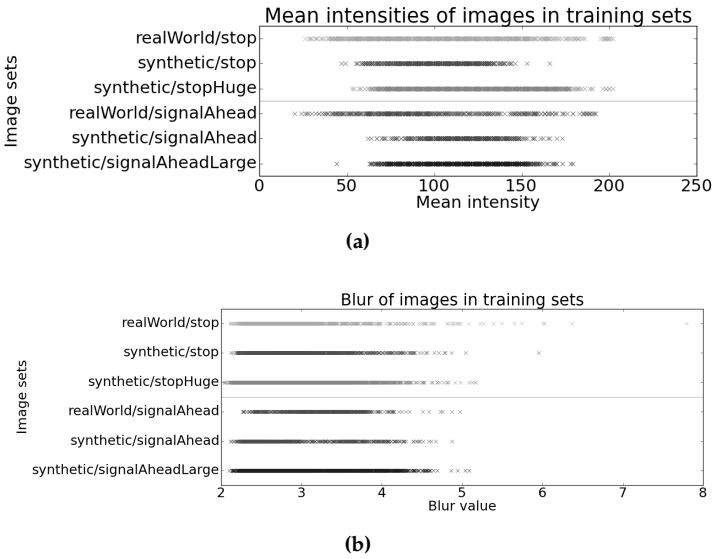


Fig. B.4: Distribution of two parameters in the training sets.

two different signs is shown. In a few sets, the intensity span does not match, but the large 5000 image stop sign set is similar to real-world data. Another parameter is shown in fig. B.4b: Blur. Blur is calculated as

$$B = \frac{1}{n} \sum_{i=0}^{n} e_i \tag{B.1}$$

where B is the blur-value, n is the number of vertical edges in an image and e_i is the edge width of a specific edge pixel, given as the distance between the pixels with the local maximum and minimum intensity around the edge pixel. The measure is described further in [7]. This shows that the blur variance is covered well by the synthetic data.

4 Comparative evaluation

To compare the synthetic training data to training data obtained from real footage, a simple Viola-Jones based detector [15] was trained for the four sign types illustrated in fig. B.2. The choice of detection algorithm is not crucial, as the purpose of this paper is not to find a perfect traffic sign detector, but rather look at the relative differences between detectors trained with synthetic and real-world images. It was trained with an image size of 20x20 pixels in all cases, except for the rectangular speed limit sign, trained with 18x24 pixels.

The detectors created with various numbers of training images was tested on a set of real-world images, collected from cars in conjunction with this lab's research. The

results can be seen in table B.1.

With all signs, the real-world data performs significantly better than the synthetic data. Providing more training data in the synthetic case does help, but even a large increase (more than a doubling) of the training data does not make the synthetic data perform comparably to the real-word data. All detectors were trained with a target of 20 stages, but some terminated earlier due to a sufficiently good fit to the training data, and others were lowered to give better detection performance at the cost of more false positives. It is indeed possible for the synthetic detector to find more true signs, but at a huge cost in false positives, and still not as good as the real-world detector.

Even in the cases (like the stop sign detector with 5000/10000 training images) where the synthetic data spans nearly the same space as the real-world detector, the synthetic detector fails to achieve a detection rate anywhere near the real-world data.

5 Concluding remarks

We discussed a research study to assess the feasibility of using carefully synthesized training datasets for developing traffic sign detectors. In this research, output from a synthetic training generator has been used to train a stock AdaBoost cascade and its performance compared with real-world training images. The real-world training data consistently performs significantly better than the synthetic training data, even in cases where the synthetic data seems to span a similar set of appearances. This leads to the conclusion that there is simply no substitute for real-world images in the case of detection.

An ML-approach to setting the synthetic data generation parameters would be a logical place to go from here, if further study of synthetic data for detection is desired. It is also possible that the system could benefit from further transformations to the template image, such as motion blur. Other works have shown promising results in using synthetic training data for classification of signs. An interesting direction of research could be to explore hybrid (real and synthetic) datasets for TSR approaches.

Bibliography

[1] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler. "A system for traffic sign detection, tracking, and recognition using color, shape, and motion information". In: *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*. IEEE. 2005, pp. 255–260.

[2] H. Fleyeh and M. Dougherty. "Road and traffic sign detection and recognition". In: *10th EWGT Meeting and 16th Mini-EURO Conference*. 2005, pp. 644–653.

[3] H. Hoessler, C. Wöhler, F. Lindner, and U. Kreßel. "Classifier training based on synthetically generated samples". In: *Proceedings of 5th international conference on computer vision systems. Bielefeld, Germany*. 2007.

[4] H. Ishida, T. Takahashi, I. Ide, Y. Mekada, and H. Murase. "Identification of degraded traffic sign symbols by a generative learning method". In: *16th International Conference on Pattern Recognition (ICPR)*. Vol. 1. IEEE. 2006, pp. 531–534.

[5] S. Lafuente-Arroyo, S. Salcedo-Sanz, S. Maldonado-Bascón, J. A. Portilla-Figueras, and R. J. López-Sastre. "A decision support system for the automatic management of keep-clear signs based on support vector machines and geographic information systems". In: *Expert Syst. Appl.* 37 (1 Jan. 2010), pp. 767–773. ISSN: 0957-4174.

[6] F. Larsson and M. Felsberg. "Using fourier descriptors and spatial models for traffic sign recognition". In: *Image Analysis* (2011), pp. 238–249.

[7] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. "A no-reference perceptual blur metric". In: *Image Processing. 2002. Proceedings. 2002 International Conference on*. Vol. 3. 2002, pp. 57–60.

[8] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets". In: *21st International Conference on Pattern Recognition (ICPR)*. IEEE, 2012, pp. 3452–3455.

[9] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey". In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (Dec. 2012), pp. 1484–1497.

[10] G. Overett, L. Tychsen-Smith, L. Petersson, N. Pettersson, and L. Andersson. "Creating robust high-throughput traffic sign detectors using centre-surround HOG statistics". In: *Machine Vision and Applications* (2011), pp. 1–14.

[11] S. Sivaraman and Mohan M. Trivedi. "A General Active Learning Framework for On-road Vehicle Recognition and Tracking". In: *IEEE Transactions on Intelligent Transportation Systems* 11.2 (June 2010), pp. 267–276.

[12] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition". In: *Neural Networks* (2012), pp. 323–332. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.02.016. URL: http://www.sciencedirect.com/science/article/pii/S0893608012000457.

[13] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. "The German Traffic Sign Recognition Benchmark: A multi-class classification competition". In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE. 2011, pp. 1453–1460.

[14] Mohan M. Trivedi and S.Y. Cheng. "Holistic sensing and active displays for intelligent driver support systems". In: *Computer* 40.5 (2007), pp. 60–68.

[15] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* 1 (2001), pp. 511–518. ISSN: 1063-6919.

Paper C

Detection of U.S. Traffic Signs

Andreas Møgelmose, Dongran Liu, and Mohan M. Trivedi

The paper has been published in the
IEEE Transactions on Intelligent Transportation Systems, in press, 2015.

The layout has been revised.

Abstract

This paper presents a comprehensive research study of the detection of US traffic signs. Until now, the research in Traffic Sign Recognition systems has been centered on European traffic signs, but signs can look very different across different parts of the world, and a system which works well in Europe may indeed not work in the US. We go over the recent advances in traffic sign detection and discuss the differences in signs across the world. Then we present a comprehensive extension to the publicly available LISA-TS traffic sign dataset, almost doubling its size, now with HD-quality footage. The extension is made with testing of tracking sign detection systems in mind, providing videos of traffic sign passes. We apply the Integral Channel Features and Aggregate Channel Features detection methods to US traffic signs and show performance numbers outperforming all previous research on US signs (while also performing similarly to the state of the art on European signs). Integral Channel Features have previously been used successfully for European signs, while Aggregate Channel Features have never been applied to the field of traffic signs. We take a look at the performance differences between the two methods and analyze how they perform on very distinctive signs, as well as white, rectangular signs, which tend to blend into their environment.

1 Introduction

Traffic sign detection has become an important topic of attention, not only for researchers in intelligent vehicles and driver assistance areas but also those active in the machine vision area. Traffic Sign Recognition (TSR) generally consists of two layers, detection and classification. With the German Traffic Sign Recognition Benchmark (GTSRB) in 2011, the classification problem was largely solved. To achieve a fully functional TSR system, the detection step needs to work as well. With the introduction of the German Traffic Sign Detection Benchmark (GTSDB) competition, a good amount of work has been done to that effect, even with suggestions of the detection problem being solved [13]. We contend that while good progress has definitely been made, the research community is not quite there yet.

Not all traffic signs look the same, especially the US signs are significantly different in appearance from those in Europe. Systems which do not consider them cannot be expected to perform in the same manner as for what they are designed for - namely almost exclusively European signs. We have taken a fresh look at the specific issues, challenges, features, and evaluation of US traffic signs in a comprehensive manner. To do this in a systematic way, the very first order of business is to draw out differences in how these signs appear. Given these rather stark appearance differences, we undertook a major database collection, annotation, organization, and public distribution effort. Secondly, we explored the overall landscape of appearance based object detection research - including European traffic signs - and carefully selected the two most promising approaches, one (Integral Channel Features) which has offered very good results on European signs and another (Aggregate Channel Features) which was very recently introduced in the literature, but has never been applied to the traffic sign case.

TSR is becoming more and more relevant, as cars obtain better and better Ad-

vanced Driver Assistance Systems (ADAS), and driving becomes more and more automated. While mapping-based indexing of traffic signs can replace in-situ recognition to some extent, it will never be able to work in changing road conditions, such as road work, and furthermore the initial sign locations and types must be determined somehow in the first place. Until the infrastructure is updated to include wireless transponders in all traffic signs, TSR will have its place in cars.

No matter the application, detecting and recognizing signs on individual images is not sufficient. If every new detection in a video feed is treated as a new sign, the driver (human or not) will quickly be overwhelmed by notifications. Instead detections must be grouped so all detections pertaining to the same physical sign are treated like the single sign it is. The temporal grouping of detections may also have a positive impact on the classification, since more than one image can be used to determine the sign type. For temporal grouping, tracking comes into play. There has been some research into tracking of traffic signs [12, 14], but it is still in its infancy. One of the issues in traffic sign tracking is that no suitable dataset exists for that purpose, something the dataset extension put forth in this paper addresses, even though we do not tackle that issue in the experiments here.

The primary goal of this paper is to present the most comprehensive treatment of the US Traffic Sign Detection studies. Specifically, this paper makes following three contributions:

- We show that while traffic sign detection has indeed come a long way over the last couple of years, international variations in traffic sign designs mean that the problem is by no means solved yet.
- We test two state-of-the-art detection schemes (ICF and ACF) on traffic signs from different countries, with special focus on US signs, which have largely been ignored by the community. We compare their results and achieve state-of-the-art detection results on both European and US signs.
- We introduce a comprehensive extension to the LISA Traffic Sign Dataset [16] which allows for detailed testing of traffic sign tracking.

The paper is structured as follows. In the next section we cover the latest related studies in the field of traffic sign detection. Section 3 briefly covers what traffic signs are, and especially how traffic sign differ among countries. We also present the extended LISA Traffic Sign Dataset. Section 4 describes which detection methods we evaluate. In section 5 we pit the detection methods against each other.

2 Related studies

Traffic sign detection has been researched seriously for about a decade. For a general overview, as well as a survey of the research up until 2012, we refer the reader to [16]. Since 2012, the efforts in detection have been stepped up. Following the successes in pedestrian detection, many of those methods have been repurposed for traffic signs. The great catalyst for the recent progress has been the German Traffic Sign Detection Benchmark (GTSDB)[8], which has really pushed the state-of-the-art detection performance to near-perfection on European signs.

Previously, the field was split in model-based and learning-based approaches, but recently the learning-based methods have taken over completely in defining the state-of-the-art. Thus, all the front-runners in the GTSDB competition were learning based. The competition encompassed 18 teams and 3 teams were considered the top performers: *Team VISICS*[13], *Team Litsi*[10], and *Team wgy@HIT501*[27]. *Team VISICS* use the Integral Channel Features (*ChnFtrs* or ICF) proposed by Dollár et. al.[6] for pedestrian detection and later improved in [3]. The same method is evaluated in this paper, along with its successor, Aggregate Channel Features. *Team Litsi* first establishes regions of interest with color classification and shape matching and the detect signs using HOG and color histograms with an SVM, features somewhat similar to ICF. Finally, *Team wgy@HIT501* uses HOG features, finding candidates with LDA and performing a more fine-grained detection using HOG with IK-SVM. In essence, all three approaches are rather similar, especially when it comes to features. Another recent paper presenting work on the GTSDB dataset is [11], which shows somewhat worse detection performance than the competitors above, but at a faster speed.

For US traffic signs, the activity has been less enthusiastic. Only a few recent studies take on US traffic signs. In 2012, [15] evaluated whether synthetic training data of US signs could be used successfully to train a rudimentary detector, but performance for the synthetic training data was poor compared to real-world images. Also in 2012, Staudenmaier et. al. [23] (building on their previous paper [24]) showed a Bayesian Classifier Cascade with intensity features and tensor features (which describe edges). They detect US speed limit signs at a good rate above 90%, but with several false positives per image, much, much worse than the current European state-of-the-art systems. Abukhait et. al. [1] use a shape-based detector to find US speed limit signs - note the model-based, rather than learning-based approach. The detector is part of a full recognition system, and the only reported performance figure is a detection rate of about 88%, but without mention of false positive rates. Stepping back in time to 2008, Keller et. al. [9] worked on detection of US speed limit signs using a voting based rectangle detector (see also [2]) followed by an AdaBoost classifier. Moutarde et. al. [17, 18] also tackled the case of US speed limit signs using a proprietary rectangle detector. Finally, a precursor to this study was presented in 2014[14]. To the best of our knowledge, no other US sign based works have been published to date. The existing papers have generally focused on speed limit signs, and achieved significantly worse performance than what we see in the GTSDB.

3 Traffic signs: International conventions and differences

The bulk of the research in TSR systems has laid in European signs, or more specifically signs conforming to the Vienna Convention on Road Signs and Signals [16, 26]. The Vienna Convention has been ratified in 62 countries, as illustrated in fig. C.1, so as an initial effort, going after those designs is reasonable. Note that while Australia and Japan have not ratified the convention, they largely design their signs in similar ways to Europe. The same holds true to a lesser extent for China. In other words, the Vienna Convention covers most of Europe and some of Asia, but leaves large parts of

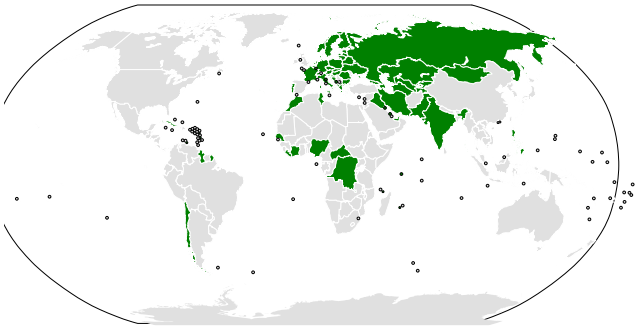


Fig. C.1: Countries which have ratified the Vienna Convention on Road Signs and Signals. Note that apart from these, Japan, Australia, and to a lesser extent China also follows it, even though they did not ratify it. Data source: [26]

Fig. C.2: Examples of Vienna Convention signs. (a) Keep right, superclass mandatory. Sign D15,3. (b) Left turn, superclass danger. Sign A41-2. (c) 60 km/h speed limit, superclass prohibitory. Sign C55.

the world out, most notably the Americas and south east Asia. Also Africa, though that continent is probably less of a market for ADAS at the moment.

The differences in sign designs matter very much in a detection context. Fig. C.2 shows a typical sign from each of the major sign superclasses in Europe: Mandatory, danger, and prohibitory. Each class is very distinctive, not only from the others, but also from most things in the real world. They all have both a rather distinctive shape and a strongly colored border/background. US signs are not in exactly the same classes, but fig. C.3 shows the matching US signs. Fig. C.4 shows more examples of US signs. From the outset, three things are obvious:

1. US signs bear little to no resemblance to Vienna Convention signs, at the very least requiring re-training of any detector.
2. The strong visual structure in Vienna Convention signs is less present in US signs. The strongest visual clue for US signs is the yellow diamond of warning signs, but even that is not present for all warning signs. The stop sign (which is identical with its Vienna Convention counterpart) is also visually strong. However, most other signs are just white rectangles of varying aspect ratios, which should be challenging to standard detectors which often rely heavily on color cues.

Fig. C.3: US signs corresponding to the Vienna Convention signs in fig. C.2. (a) Keep right, superclass traffic movement. Sign R4-7. (b) Left turn, superclass warning. Sign W1-1. (c) 50 mph speed limit, superclass speed limit. Sign R2-1. Image source: [22]

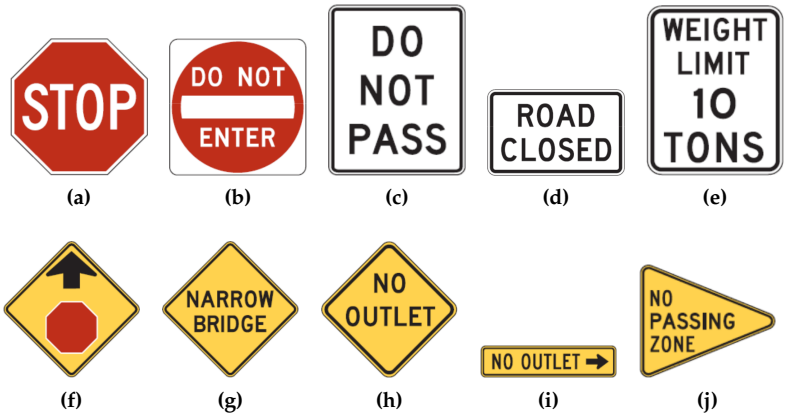


Fig. C.4: Examples of US signs. (a) Sign R1-1. (b) Sign R5-1. (c) Sign R4-1. (d) Sign R1-2. (e) Sign R12-1. (f) Sign W3-1. (g) Sign W5-2. (h) Sign W14-2. (i) Sign W14-2a. (j) Sign W14-3. Image source: [22]

3. Many US signs contain only text to convey their message, as opposed to Vienna Convention signs which mostly use icons and pictograms.

Given these large differences to Vienna Convention signs, and the size of the car market in the US, it is surprising that in [16], only two studies were concerned with US traffic signs, and as we describe in the previous section on related studies, not many have come out since the publication of the review.

3.1 Dataset

In order to properly train and evaluate TSR systems, traffic sign datasets are needed. In [16], the LISA Traffic Sign Dataset was introduced. In this paper we announce a very large extension of the LISA Traffic Sign Dataset, which almost doubles its size with the addition of annotated high-resolution color images. The LISA-TS Extension is split into a training set with every 5th frame annotated and a test set with every frame annotated, so it is also suitable for testing traffic sign tracking systems. Tracking

is outside of the scope of this paper, but see [14] for a simple tracking experiment and [12] for a more advanced solution.

The extension has been collected in and around San Diego, California on urban streets during the spring of 2014. All images were captured with a Point Grey FL3 color camera at the resolution of 1280x960 at approximately 15 fps. Weather conditions are generally dry and sunny or cloudy.

Table C.1: Traffic Sign Dataset statistics

	Original LISA Traffic Sign Dataset	**LISA-TS Extension**			**GTSD[8]**	**BTSD[13]**
		Training set	Testing set	Combined		
Number of classes:	47	31	17	35	4	4
Number of annotations:	7855	3672	2422	6094	1206	4627
Number of images:	6610	3307	2331	5638	900	9006
Sign sizes, longest edge:	6 – 168 px	23 – 222 px	25 – 222 px	23 – 222 px	16 – 128 px	16 – 913 px
Image sizes:	640x480 to 1024x522 px	1280x960 px	1280x960 px	1280x960 px	1360x800 px	1628x1236 px
Includes videos:	Yes, but not public	Yes	Yes	Yes	No	Image sequences
Video annotation density:	Every 5 frames	Every 5 frames	All frames	Mixed	N/A	N/A
Notes	Some images in grayscale				The provided classes are actually superclasses.	Classes are superclasses. Multi-view: 8 cameras on one car.

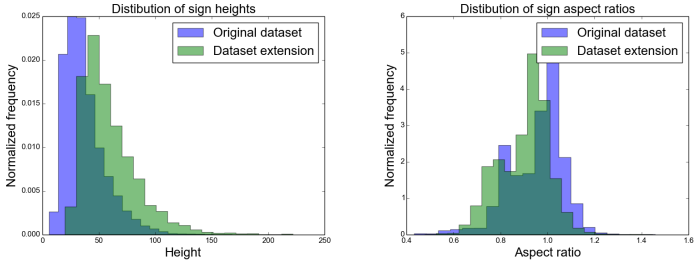


Fig. C.5: Histograms of the (a) heights and (b) aspect ratios of the annotated signs. Statistics for the original dataset and for the extension have been overlaid for easy comparison.

Table C.1 shows statistics about both LISA-TS and LISA-TS Extension, and compares them to the two other relevant detection datasets, the German Traffic Sign Detection set (GTSD)[8] and the Belgium Traffic Sign Detection set (BTSD)[13]. When LISA-TS and LISA-TS Extension are combined, the dataset size exceeds even BTSD.

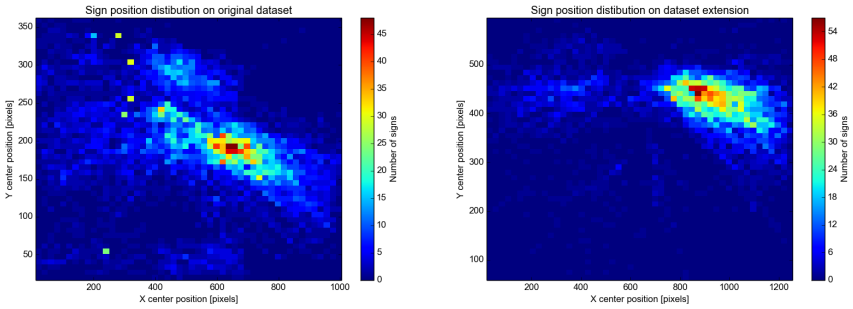
Fig. C.5 shows the size and aspect ratio distributions for the original LISA-TS set and the LISA-TS Extension. Height-wise we see a similar distribution on both datasets, but shifted towards larger sizes for the extension. This fits well with the extension being higher resolution (sizes are measured in pixels) and the similar distributions show that both datasets cover signs at approximately the same distances. With regards to aspect ratio, the distributions are also similar. Many signs have an aspect ratio of 1.0 - this covers round, square and octagonal signs, and the remaining are around 0.8, which fits well with speed limit signs and other rectangular signs. Ratios outside of this is explained by non-orthogonal viewing angles, which can significantly distort the sign in the image plane.

Position heatmaps are shown in fig. C.6, one for the original set and one for the extension. In both cases most signs are clearly positioned along the right shoulder of the road, as expected. The original set is somewhat less clearly defined than the extension, undoubtedly because the original set was captured from different vehicles with slightly differing camera positions, whereas the extension has been captured with a single vehicle with a fixed camera position.

Qualitatively, there are some differences between LISA-TS and LISA-TS Extension. LISA-TS predominantly consists of low resolution images, some in grayscale, but was captured over a larger area in Southern California. LISA-TS Extension has consistent high-res color images, all captured from the same rig, ensuring similar images across the set. It also has a dedicated test set. The Extension was captured in and around San Diego. Both datasets suffer from the magnificent sunny Californian weather, so researchers interested in evaluating their algorithms in adverse weather conditions should look elsewhere.

LISA-TS is used as the benchmarking dataset for the sign-detection part of the VIVA 2015 workshop[1] at Intelligent Vehicles Symposium, 2015. For a further overview

[1] http://cvrr.ucsd.edu/vivachallenge/



(a) **(b)**

Fig. C.6: Heatmaps showing the positions of annotations in the frame for (a) the original dataset and (b) the extension. The contours of the road can almost be seen in the plots, especially in the extension, which has been captured using an identical camera setup for all frames. The heat maps give a very strong hint towards reasonable Regions of Interest for traffic sign detectors.

of other datasets, see [16]. In this paper we use both the GTSD and the expanded LISA-TS dataset.

4 Detection methods

We evaluate two state-of-the-art detection methods on traffic sign detection: Integral Channel Features and Aggregate Channel Features. Both are adapted from pedestrian detection and the first has previously been used for traffic sign detection with great success[13]. Our implementation takes its starting point in the Matlab code of Piotr Dollár's Computer Vision Matlab Toolbox[5], with the settings based on his as well. We also discuss image pre-processing, as we find that color normalization is absolutely crucial for good detection performance. Figure C.7 shows the flow of a TSR system. Some systems may incorporate just a subset of the blocks shown, but we have included all to give the reader an understanding of the full process. In this paper we focus on capturing (via the dataset capture), pre-processing, feature extraction, and detection. Tracking is not part of the experiments presented here.

4.1 Image pre-processing

We use contrast-limited adaptive histogram equalization (CLAHE)[28] to normalize colors in the input images. It is a type of histogram equalization which works on tiles in the image, in order to reduce the excessive contrast and noise that may arise from ordinary histogram equalization. Ordinary histogram equalization is done by mapping pixels to a different value based on the cumulative distribution function (CDF) of pixel values in the image. Adaptive Histogram Equalization (AHE), which is slightly simpler than CLAHE, works by performing this transformation for each pixel only by considering the CDF of pixels nearby - a tile. This means that contrast is

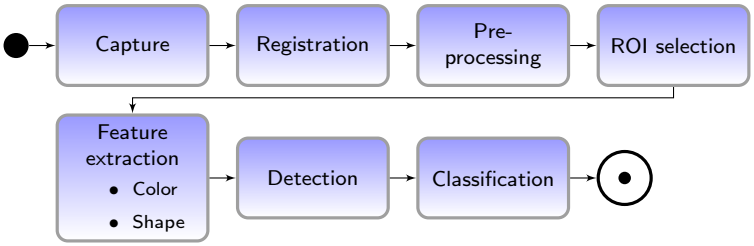


Fig. C.7: Flowchart of a typical TSR system. A picture is captured and if the systems supports detailed real-world positioning of signs, a registration of this image happens. Afterwards, some pre-processing takes place, usually color normalization followed by an ROI-selection if necessary. This can speed up detection by only looking in relevant parts of the image. Usually the ROI is hardcoded in advance, but saliency measures may also be used. Then relevant features are extracted, most often from a sliding window, detection happens, and finally the detected signs are classified. Some systems contain only some of these blocks, and some systems have this pipeline running in parallel for different sign superclasses.

locally enhanced. A potential problem arises with AHE, though. If a tile is uniform, its CDF has a strong peak, which amplifies pixel noise. CLAHE attempts to combat this by clipping the peak of the CDF so no pixels are too heavily amplified. To speed up computation, the equalization is performed in non-overlapping tiles, which are then blended using bilinear interpolation. Fig. C.8 shows the precision-recall curve for stop sign detection with and without CLAHE, clearly demonstrating the importance of this step.

4.2 Integral Channel Features

Integral Channel Features (ICF, sometimes also abbreviated as ChnFtrs) was first presented by Piotr Dollár for pedestrian detection [6], and recently repurposed by Mathias et. al. [13] to achieve near-perfect detection on GTSD. In this paper we put this algorithm to the test on US traffic signs.

The method has two key ingredients: Features computed over 10 different "channels" of the image, and detection using a boosted decision forest. First, the input image is split into 10 channels. The channels used are the three LUV color channels, one for unoriented gradient magnitude, and six for gradients in varying directions. For each of these channels, first-order Haar-like features are computed. This is simply the difference of the sum of two different rectangular image regions. While higher-order features are a possibility, the gain from those is very low. Feature computation is sped up by using the integral image of each channel, C, defined as

$$CC_j(x,y) = \sum_{x' \leq x, y' \leq y} C_j(x',y'), j = 1, \ldots, D \tag{C.1}$$

where D is the number of channels. In [13], detectors for each sign are trained for various skewed aspect ratios, to account for non-orthogonal viewing angles. We found only negligible performance gains from this and thus do not have that as part of our training.

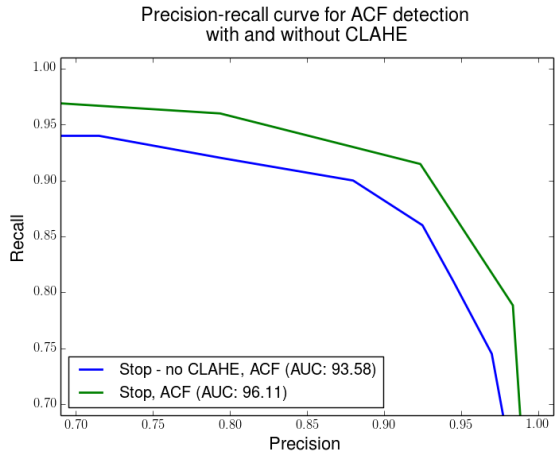


Fig. C.8: Precision-recall curves for detection of stop signs with and without color pre-processing. The advantage of color pre-processing is evident across the curve.

After the features are computed, an AdaBoost classifier is learned with depth-2 decision trees as weak learners. This classifier is then run on a sliding window on the input image.

4.3 Aggregate Channel Features

In 2014, Dollár et. al. published an enhanced version of ICF, called Aggregate Channel Features (ACF)[4]. Ostensibly, ACF was introduced as a faster alternative to ICF, but in some cases it shows better detection performance as well. The basic principle about computing features across channels is the same as ICF, and indeed the same channels are used. The Haar-like features are replaced with an even simpler scheme: summing up blocks of pixels at various scales. This is obviously faster than computing the already simple Haar features, but as we shall see provides similar and sometimes even better detection. The boosted decision forest is preserved as the classifier of choice.

5 Evaluations

We evaluate the two detectors on both the GTSD (to verify that we can replicate the near-perfect results of [13]) and more importantly on the LISA-TS, to show how the methods work on US signs and highlight the challenges unique to US traffic signs. The PASCAL measure [7] has been used to determine detection rates, as is standard. A detection is considered true if the overlap of the detection bounding box and the annotation is more than 50%:

$$a_o \equiv \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{C.2}$$

Table C.2: LISA-TS training and testing statistics

	Number of images	
Superclass	**Training**	**Test**
Diamond	1229	406
Stop	1182	1152
NoTurn	185	83
SpeedLimit	750	680

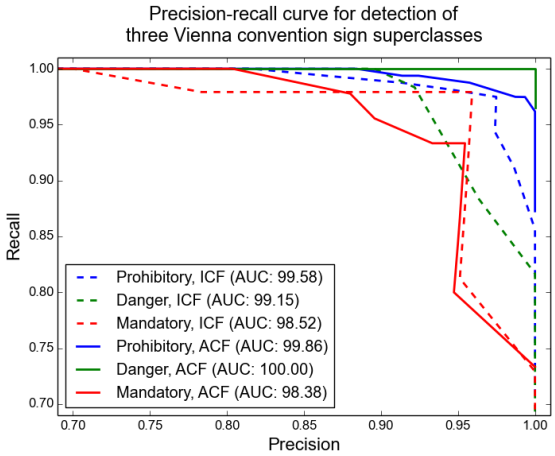


Fig. C.9: Precision-recall curves for detection of Vienna convention mandatory signs, prohibitory signs, and danger signs. Both ICF and ACF are shown for each superclass. Note that the axes are zoomed to provide a more detailed picture.

where BB_{dt} is the bounding box of the detection and BB_{gt} the bounding box of the ground truth.

5.1 European signs

The GTSB is divided into a separate training and test set and spans 4 superclasses: *prohibitory* signs (circular with a red border), *mandatory* signs (circular and blue), *danger* signs (triangular with a red border), and *other* signs, comprising all signs which do not fit in any of the three other categories. *Other* is a very diverse category spanning many shapes and colors, and since it was not considered in the GTSD benchmark, we also ignore it.

Results for each of the three superclasses are shown in fig. C.9. ACF detects danger signs perfectly - no misses and no false positives. The detection is close to perfect for the remaining two classes as well. Overall, the performance is comparable to that of Mathias et. al. in [13]. We fare a little worse on prohibitory signs at an

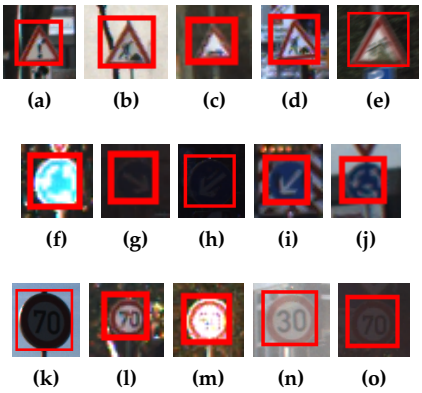


Fig. C.10: Examples of Vienna convention signs which ICF misses. By row: Danger, mandatory and prohibitory.

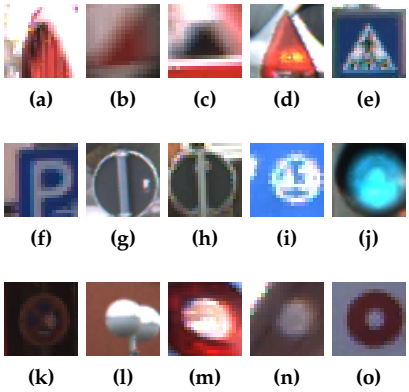


Fig. C.11: Examples of false positives on Vienna convention signs for ICF. By row: Danger, mandatory and prohibitory.

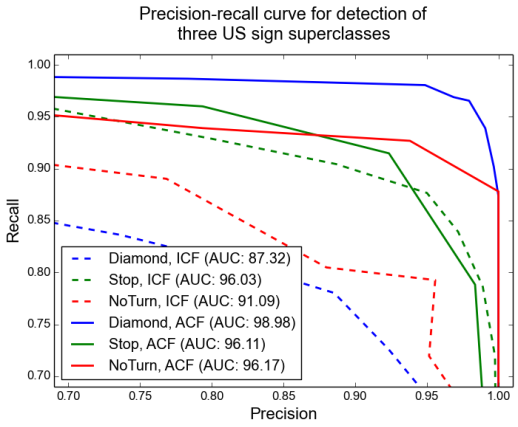


Fig. C.12: Precision-recall curves for detection of US stop signs, warning signs, and no-turn signs. Both ICF and ACF are shown for each superclass. Note that the axes are zoomed to provide a more detailed picture.

Area Under Curve (AUC) of 99.58/99.86 for ICF/ACF vs. their 100 and a little better on mandatory signs with 98.52/98.38 vs. their 96.98. ICF generally performs very slightly lower than ACF, due to just a few troublesome images, see figures C.10 and C.11. It is not impossible that both could be further tweaked to obtain perfect scores.

5.2 US signs

We tested the algorithms on the extended LISA-TS dataset, split up into a training set and a test set. In total, the training set comprise 3346 pictures and the test set contains 2321 pictures. Details for each superclass can be seen in table C.2.

Detection results for US signs are shown in fig. C.12. The precision-recall is generally worse than for European signs, though the diamond superclass just barely surpasses the European mandatory superclass. This shows that even the best performing methods for European signs do not necessarily generalize to other traffic sign design schemes. It is also clear that ACF performs significantly better than ICF for US signs, whereas the difference for European signs was much less pronounced. Diving into the numbers, diamond signs are still detected with what can be considered very good performance with an AUC of 98.98 for ACF. The two other superclasses also have an AUC above 95, but there is room for improvement. Examples of false detections and misses can be seen in fig. C.13 and C.14.

5.3 US speed limit signs

As shown above, ICF and especially ACF performs very well on European signs and "easy" US superclasses. Common for all these signs is that they have very strong color and shape cues. But a large amount of US signs are not that easy to distinguish (see section 3). Many are simply white rectangles. A good representative of this design is

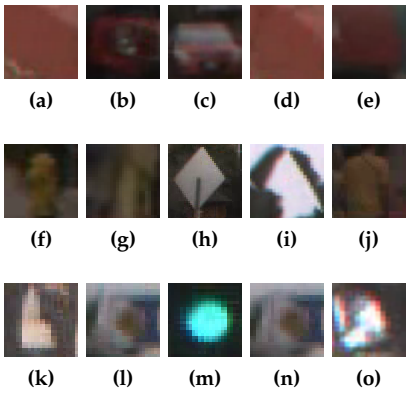


Fig. C.13: Examples of false positive on US signs. Each row is one superclass: Stop, diamond, and no turn.

the speed limit superclass, shown in fig. C.3c. Thus, we dedicate this section to that specific superclass.

We have run both our ICF and ACF detector on this superclass, and indeed the performance is worse. Fig. C.15 shows precision-recall curves for both detectors. The other US superclasses have been added for reference. The performance is significantly worse than for the more salient superclasses with AUCs under 90 for both detectors. Still, we note that the performance is much better than the competing US sign detection systems mention in Related Studies (section 2). Interestingly, ICF performs slightly better than ACF for this particular superclass. Since there is no color in these signs, the LUV channels used by both detection schemes have very limited impact, other than discarding brightly colored objects. It is also unlikely that a better color normalization scheme will have a significant impact, as was the case for colored signs. Examples of false detections and misses can be seen in fig. C.16.

We consider detection of this superclass and its colorless rectangular siblings an open problem. Note also that in this study, we have only looked at speed limit signs, and there is a whole host of other very similar designs. It is not clear from this study whether they should all be lumped together in a monolithic detector, or it is better to have multiple dedicated detectors. While there can be significantly semantic difference for humans between a speed limit sign and a do not pass sign, they have a very similar design, pointing towards the monolithic solution. On the other hand, speed limit signs have very large characters for the numbers, a visual contrast to other text based signs. Furthermore, some signs like that in fig. C.3a contain no text at all. It is also worth considering that many of the white rectangular signs have different aspect ratios.

5.4 Channel analysis

To better understand the contributions of individual components in detection for different signs, we have run detectors using color channels only and shape channels

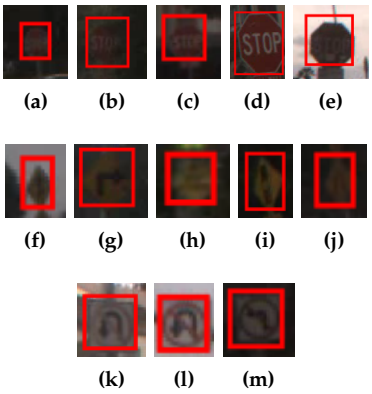


Fig. C.14: Examples of missed US signs. Each row is one superclass: Stop, diamond, and no turn.

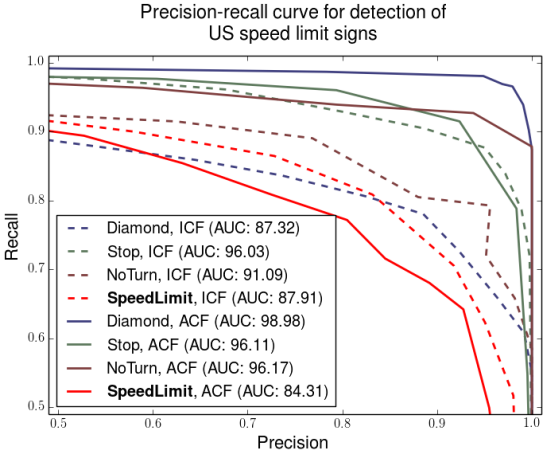


Fig. C.15: Precision-recall curves for detection of US speed limit signs. For comparison, stop signs, warning signs, and no-turn signs are also shown. Both ICF and ACF are shown for each superclass. The axes are zoomed to provide a more detailed picture, but not as much as for previous figures.

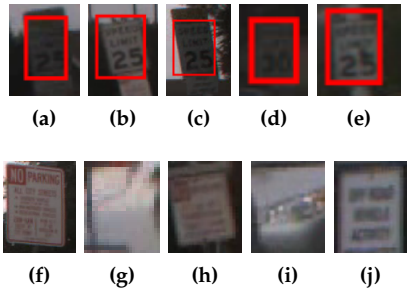
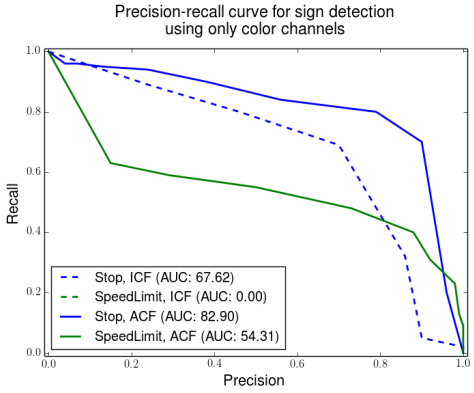
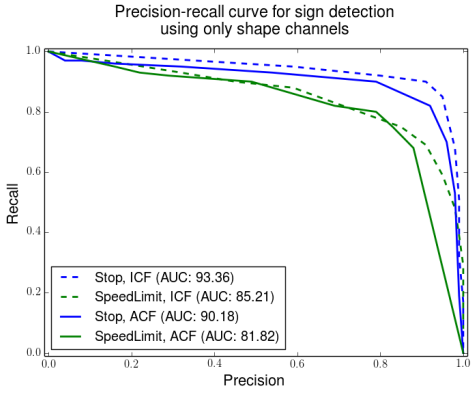


Fig. C.16: Examples of troublesome US speed limit signs. First row shows missed signs and second row shows false positives.



(a)



(b)

Fig. C.17: Detection performance using a) only color channels and b) only shape channels for stop and speed limit signs. Shape is a stronger cue.

only (fig. C.17). The broad conclusion is that shape is a stronger cue than color. In particular for ICF, where the shape-only detection is close to the performance with all channels for both superclasses - ICF also fails completely on speed limit signs when only color is used. ACF shows a stronger reliance on color cues, though shape is still the most important. This helps explain why ICF performs better than ACF for speed limit signs when all channels are used. Overall, it is interesting to see how differently color and shape contribute in the two methods, given that both methods are building on the same principles. In all cases, though, the combined detector performs better than either of the channel subsets on their own.

Unsurprisingly, color detection works much better on stop signs which have a strong red color, than on speed limit signs, which are just white so the color cue can only be used to discard strongly colored candidate windows.

6 Concluding remarks

Traffic sign detection has come far on European signs, but not much attention has been given to US signs. This study remedies this discrepancy, by bringing detection of some US sign types - diamond warning signs, stop signs, and no turn signs - up on par with detection rates for European traffic signs with AUCs of 98.98, 96.11. and 96.17, respectively. We test the established ICF detector on US signs and are the first to bring the newer and (in some cases) better ACF algorithm to the domain of traffic sign detection.

Our analysis shows that while we achieve better detection rates on speed limit signs than any studies before us, there is still work to be done on that particular superclass. It is not a given that the methods tested in this paper are the way to go when detecting signs with limited color- and shape cues. Furthermore we provide a large extension to the existing LISA-TS traffic sign dataset, which is publicly available and the only large-scale dataset of US traffic signs in existence.

The most obvious place to direct future research - at least in pure detection - is to push the boundary in detection of US speed limit signs and all the visually related white rectangular signs. Of course tracking is also very relevant for combining distinct detections of the same sign into a single entity, but as indicated by [14], very complicated tracking schemes are perhaps unnecessary in the TSR domain.

A detection method based off of the same methods used here was presented in [19], and it might be possible to adapt to traffic signs and account for detection of traffic signs which are significantly skewed with respect to the image plane. Another interesting aspect could be to combine traffic sign detection with vehicle detection [21] for a more holistic understanding of the traffic situation - for example when a sign signals a lane merge, the position of other cars is very relevant.

Finally, combining a TSR system with a driver attention estimation system, such as in [25], could result in a driver assistance system which dynamically informs the driver only about relevant signs he or she has not seen. This could significantly reduce the rate of signs missed by the driver, while also not causing information overload - as shown in [20] there are whole classes of signs which drivers are very bad at noticing, so such a system would certainly be valuable.

1 LISA US Traffic Sign Data Set: Expanded Version

The LISA Traffic Sign Dataset is the first and only publicly available data set containing US traffic signs. It was introduced in the 2012 paper *Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey* by Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund [16].

It is intended to be used in the development of Traffic Sign Recognition (TSR) systems. The original dataset was captured and released in 2012. In 2014, an extension was made to include higher resolution color images and a dedicated test set where every frame is annotated, to facilitate evaluation of traffic sign tracking systems.

LISA-TS can be downloaded from `http://cvrr.ucsd.edu/LISA/lisa-traffic-sign-dataset.html`.

Main highlights of the contents and classes of the LISA-TS data set are presented below.

1.1 Breakdown by class

Original dataset

Superclasses			
3314	warning	1508	speed limit
73	noTurn		
Detailed classes			
294	addedLane	34	slow
37	curveLeft	11	speedLimit15
50	curveRight	349	speedLimit25
35	dip	140	speedLimit30
23	doNotEnter	538	speedLimit35
9	doNotPass	73	speedLimit40
2	intersection	141	speedLimit45
331	keepRight	48	speedLimit50
210	laneEnds	2	speedLimit55
266	merge	74	speedLimit65
47	noLeftTurn	132	speedLimitUrdbl
26	noRightTurn	1821	stop
1085	pedestrianCrossing	168	stopAhead
11	rampSpeedAdvisory20	5	thruMergeLeft
5	rampSpeedAdvisory35	7	thruMergeRight
3	rampSpeedAdvisory40	19	thruTrafficMergeLeft
29	rampSpeedAdvisory45	60	truckSpeedLimit55
16	rampSpeedAdvisory50	32	turnLeft
3	rampSpeedAdvisoryUrdbl	92	turnRight
77	rightLaneMustTurn	236	yield
53	roundabout	57	yieldAhead
133	school	21	zoneAhead25
105	schoolSpeedLimit25	20	zoneAhead45
925	signalAhead		

In total: 7855 sign annotations

Dataset extension: Training

Superclasses			
1232	warning	752	speed limit
184	noTurn		
Detailed classes			
2	addedLane	91	school
20	bicyclesMayUseFullLane	428	signalAhead
50	curveLeft	4	speedBumpsAhead
59	curveRight	17	speedLimit15
39	doNotEnter	259	speedLimit25
11	intersection	90	speedLimit30
24	intersectionLaneControl	158	speedLimit35
127	keepRight	92	speedLimit40
47	laneEnds	80	speedLimit45
6	leftAndUTurnControl	53	speedLimit50
18	merge	3	speedLimit60
73	noLeftAndUTurn	1181	stop
8	noParking	86	stopAhead
16	noRightTurn	4	yieldAhead
95	noUTurn	8	yieldToPedestrian
523	pedestrianCrossing		

In total: 3672 sign annotations

Dataset extension: Testing

Superclasses			
461	warning	679	speed limit
82	noTurn		
Detailed classes			
13	curveRight	40	signalAhead
17	dip	406	speedLimit25
21	doNotEnter	264	speedLimit30
11	keepRight	9	speedLimit45
37	merge	1151	stop
29	noLeftTurn	86	stopAhead
53	noUTurn	43	turnRight
80	pedestrianCrossing	145	warningUrdbl
17	school		

In total: 2422 sign annotations

1.2 List of superclasses

warning	addedLane	**speedLimit**	speedLimit15
	curveLeft		speedLimit25
	curveRight		speedLimit30
	dip		speedLimit35
	intersection		speedLimit40
	laneEnds		speedLimit45
	merge		speedLimit50
	pedestrianCrossing		speedLimit55
	roundAbout		speedLimit60
	signalAhead		speedLimit65
	slow		
	speedBumpsAhead		
	stopAhead		
	thruMergeLeft		
	thruMergeRight		
	turnLeft		
	turnRight		
	yieldAhead		
	warningUrdbl		
noTurn	noLeftAndUTurn		
	noUTurn		
	noLeftTurn		
	noRightTurn		

Acknowledgment

The authors would like to thank their colleagues at the LISA lab for useful discussion and encouragement, especially Eshed Ohn-Bar for his valuable comments. We would also like to thank the reviewers for their thoughtful comments.

Bibliography

[1] Jafar Abukhait, Imad Zyout, and Ayman M Mansour. "Speed Sign Recognition using Shape-based Features". In: *International Journal of Computer Applications* 84.15 (2013), pp. 31–37.

[2] N. Barnes and G. Loy. "Real-time regular polygonal sign detection". In: *Field and Service Robotics*. Springer. 2006, pp. 55–66.

[3] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. "Pedestrian detection at 100 frames per second". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2903–2910.

[4] P. Dollar, R. Appel, S. Belongie, and P. Perona. "Fast Feature Pyramids for Object Detection". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.8 (Aug. 2014), pp. 1532–1545. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2300479.

[5] Piotr Dollár. *Piotr's Computer Vision Matlab Toolbox (PMT)*. http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[6] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. "Integral Channel Features". In: *BMVC*. Vol. 2. 3. 2009, p. 5.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.

[8] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark". In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE. 2013, pp. 1–8.

[9] C.G. Keller, C. Sprunk, C. Bahlmann, J. Giebel, and G. Baratoff. "Real-time recognition of U.S. speed signs". In: *Intelligent Vehicles Symposium, IEEE*. June 2008, pp. 518–523. DOI: 10.1109/IVS.2008.4621282.

[10] Ming Liang, Mingyi Yuan, Xiaolin Hu, Jianmin Li, and Huaping Liu. "Traffic sign detection by ROI extraction and histogram features-based recognition". In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. Aug. 2013, pp. 1–8. DOI: 10.1109/IJCNN.2013.6706810.

[11] Chunsheng Liu, Faliang Chang, and Zhenxue Chen. "Rapid Multiclass Traffic Sign Detection in High-Resolution Images". In: *Intelligent Transportation Systems, IEEE Transactions on* 15.6 (Dec. 2014), pp. 2394–2403. ISSN: 1524-9050. DOI: 10.1109/TITS.2014.2314711.

[12] M. Boumediene, J.-P. Lauffenberger, J. Daniel, and C. Cudel. "Coupled Detection, Association and Tracking for Traffic Sign Recognition". In: *Intelligent Vehicles Symposium (IV), 2014 IEEE*. June 2014, pp. 1402–1407.

[13] Markus Mathias, Radu Timofte, Rodrigo Benenson, and Luc Van Gool. "Traffic sign recognition — How far are we from the solution?" In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE. 2013, pp. 1–8.

[14] Andreas Møgelmose, Dongran Liu, and Mohan M. Trivedi. "Traffic Sign Detection for U.S. Roads: Remaining Challenges and a Case for Tracking". In: *17th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2014, pp. 1394–1399.

[15] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets". In: *21st International Conference on Pattern Recognition (ICPR)*. IEEE, 2012, pp. 3452–3455.

[16] Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund. "Vision based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey". In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (Dec. 2012), pp. 1484–1497.

[17] F. Moutarde, A. Bargeton, A. Herbin, and L. Chanussot. "Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system". In: *Intelligent Vehicles Symposium*. IEEE. 2007, pp. 1122–1126.

[18] Fabien Moutarde, Alexandre Bargeton, Anne Herbin, and Lowik Chanussot. "Modular Traffic Sign Recognition applied to on-vehicle real-time visual detection of American and European speed limit signs". In: vol. 14. 2009.

[19] Eshed Ohn-Bar and Mohan M. Trivedi. "Fast and robust object detection using visual subcategories". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014, pp. 179–184.

[20] David Shinar. *Traffic safety and human behaviour*. Emerald Group Publishing, 2007. ISBN: 978-0-08-045029-2.

[21] S. Sivaraman and Mohan M. Trivedi. "Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis". In: *Intelligent Transportation Systems, IEEE Transactions on* 14.4 (Dec. 2013), pp. 1773–1795.

[22] State of California, Department of Transportation. *California Manual on Uniform Traffic Control Devices for Streets and Highways*.

[23] Armin Staudenmaier, Ulrich Klauck, Ulrich Kreßel, Frank Lindner, and Christian Wöhler. "Confidence Measurements for Adaptive Bayes Decision Classifier Cascades and Their Application to US Speed Limit Detection". In: *Pattern Recognition*. Vol. 7476. Lecture Notes in Computer Science. 2012, pp. 478–487.

[24] *Resource Optimized Cascaded Perceptron Classifiers using Structure Tensor Features for US Speed Limit Detection*. Vol. 12. 2011.

[25] Ashish Tawari, Andreas Møgelmose, Sujitha Martin, Thomas B. Moeslund, and Mohan M. Trivedi. "Attention Estimation by Simultaneous Analysis of Viewer and View". In: *17th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Oct. 2014, pp. 1381–1387.

[26] United Nations. *Vienna Convention on Road Signs and Signals*. 1978.

[27] Gangyi Wang, Guanghui Ren, Zhilu Wu, Yaqin Zhao, and Lihui Jiang. "A robust, coarse-to-fine traffic sign detection method". In: *Neural Networks (IJCNN), The 2013 International Joint Conference on*. Aug. 2013, pp. 1–5. DOI: 10.1109/IJCNN.2013.6706812.

[28] Karel Zuiderveld. "Contrast limited adaptive histogram equalization". In: *Graphics gems IV*. Academic Press Professional, Inc. 1994, pp. 474–485.

Part III

Pedestrian Detection and Analysis

Paper D

Part-based Pedestrian Detection and Feature-based Tracking for Driver Assistance: Real-Time, Robust Algorithms and Evaluation

Antonio Prioletti, Andreas Møgelmose, Paolo Grisleri, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund

The paper has been published in the
IEEE Transactions on Intelligent Transportation Systems Vol. 14.3,
pp. 1346–1359, 2013.

The layout has been revised.

Abstract

Detecting pedestrians is still a challenging task for automotive vision system due the extreme variability of targets, lighting conditions, occlusions, and high speed vehicle motion. A lot of research has been focused on this problem in the last 10 years and detectors based on classifiers has gained a special place among the different approaches presented. This work presents a state-of-the-art pedestrian detection system based on a two stages classifier. Candidates are extracted with a Haar cascade classifier trained with the DaimlerDB dataset and then validated through part-based HOG classifier with the aim of lowering the number of false positives. The surviving candidates are then filtered with a feature-based tracking to enhance the recognition robustness and improve the results' stability. The system has been implemented on a prototype vehicle and offers high performance in terms of several metrics, such as detection rate, false positives per hour, and frame rate. The novelty of this system rely in the combination of HOG part-based approach, tracking based on specific optimized feature and porting on a real prototype.

1 Introduction

Pedestrian detection is a popular field in research on intelligent vehicles, but even though it has been met with intense research, it is still not a solved problem. According to [9], it is even unlikely that an acceptable detection rate will be reached "without major leaps in our understanding." Pedestrian detection has multiple uses, the most prominent being Advanced Driver Assistance Systems (ADAS). The overarching goal is to equip vehicles with sensing capabilities to detect and act on pedestrians in dangerous situations, where the driver would not be able to avoid a collision. A full ADAS with regards to pedestrians would as such not only include detection, but also tracking, orientation, intent analysis, and collision prediction.

Pedestrian detection brings many challenges, as outlined by [23]: high variability in appearance among pedestrians, cluttered backgrounds, high dynamic scenes with both pedestrian and camera motion, and strict requirements in both speed and reliability. It follows from this list that there is a high risk of occlusions, but also that those occlusions might not be present for very long, since all objects in the scene are moving relatively to each other. Part-based detection systems seem intuitive to cope well with occlusions, as they do not necessarily require the full body to be present to make a detection. Also, many existing systems (see Section 2) are plagued by a high false positive per frame, something that a part-based system can reduce if requirements of several body-parts to be detected are put in place. These two motivations for part-based detection can be somewhat contradictory: narrowing the classification parameters will reduce the number of false positives but, likewise, the number of true positives. A tracking technique can be introduced to supply missing detections and thus counteract this trade-off.

This paper builds on a part-based, staged detection approach, which was first put forth in [32], providing 4 major contributions:

1. a thorough analysis of the impact of changes in parameters for this algorithm that goes far beyond what was presented in the initial study.

2. An expansion of the system to a full-fledged ADAS, not just a detection algorithm, and a discussion of the requirements put upon the full system from such an application.
3. The use of more pedestrian related training and test sets, where the original paper used the INRIA dataset [7], which is a more general purpose person dataset.
4. Porting of the system to a real prototype vehicle and analysis of critical situations in a real environment, optimizing the system to improve detections and speed performance.

One of the innovations of this system is the usage of HOG features in a part-based approach; moreover an optimized kind of feature has been adopted to decrease as much as possible the computational time this helps when test the system on a real prototype. Given the reaction speed of a human, it is clear that a braking assistance system can help reducing braking distances.

ADAS is a challenging domain to work within. Reaction times must be fast for driving, where a fraction of a second can be the deciding factor between a collision and a near-miss. At the same time, the system must be robust, so no action is erroneously triggered (due to a false detection), which could itself lead to accidents. Further reasoning than just detection is necessary in such a framework, with pedestrian intent estimation being a good example, as presented in [26] or automatic braking as in [5].

This paper contains an overview of related works in section 2, a description of the implemented pedestrian detection ADAS in section 3, and details of the algorithmic stages in sections 3.1, 3.2, and 3.3. A thorough set of experiments follows in section 4 where the impact of parameter adjustments in the system are investigated. Section 5 describes the porting of the system to a real prototype car and section 6 presents a final evaluation of the performance, compared to the state-of-the-art of vision based detectors, with the full-body approach by Geismann et. al. and with the final system after the implementation on a real platform. [22].

2 Related works

The purpose of pedestrian detection is first and foremost to protect pedestrians. Pedestrian safety is a large area including passive solutions, such as car design, and active solutions, such as pedestrian detection. It also involves infrastructure design to a great extent. [18] provides a survey of the pedestrian detection field and a taxonomy of the involved system types. Many standard features and learning algorithms have been adapted to pedestrian detection. Common options include an AdaBoost cascade on Haar-like features [41, 24] or HOG+SVM [7, 46], but many other features are also used, such as edgelets [43], variations of gradient maps, or simple intensity images. The cascade classifier based on Haar-like features, described by Viola and Jones [41], is a very fast algorithm for pedestrian detection. A drawback of this approach is the close link with the appearance of pedestrians and the resulting lack of robustness. An alternative is the HOG-SVM solution presented by Dalal and Triggs [7]. At the cost of speed, this algorithm is much more robust and detects pedestrians in harder

situations. The combination of these two algorithms allows the system to benefit from both approaches and obtain a robust system with considerable speed-up.
Decompose the pedestrian shape into parts is gaining great interest in this area, particularly for increased tolerance of occlusions. Interesting dilemmas are how many and which parts of pedestrians to use, and how to integrate all the part-based detectors in a final detector; an example is shown in [33] where in the first stage head, arm and leg detectors were trained in a fully supervised manner and then combined the detector to fit a rough geometric model. Other two stage approaches are shown in [31, 11]. Several feature types and different environment kinds can be used. Mao et. al [30] developed a system based on Viola's Adaboost cascade framework, using edgelets features in addition to Haar-like features to improve the detection of the pedestrians contour; moreover it was introduce the concept of interfering object: object similar to human body on a feature level. Before detect pedestrians they remove this type of object. In [43], Nevatia et. al, combine multiple part detectors based on edgelet to form a joint likelihood model that includes cases of multiple, possibly inter-occluded humans. Due to the high difficulty of detecting interest point at low resolutions, unsupervised part based approaches that do not rely on keypoints have been proposed. An example is Multiple Instance Learning (MIL), that determine the position of parts without part-level supervision [47]. Felzenszwalb et al. [16] proposed one of the most successful part-based approaches that models unknown part positions as latent variables in a SVM framework. [36] improves this method switching to part-based system only at sufficiently high resolutions. Detecting a highly variable objects such as pedestrians is essential the use of a tracking module. Tracking a variable number of elements in complex scenes is a challenging process. To cope with this kind of problems is becoming common [3, 43] use a tracking-by-detection approach: detect pedestrians in individual frames and associate them between frames. The main challenge regards a discontinuous detection in conjunction with a possible false positives and missing detection; this problem makes hard to use a Kalman filter, due to the continuous detection that it needs to give accurate results. Several multi-object tracking systems [2, 27], as our system, use a large temporal window to make the association; in this way a pedestrian not detected in two subsequent frames but in more frames can be also included in the tracking system with a temporal delay. An other interesting approach, that can be investigated in the future, is to represent the uncertainty of a tracking system with a Particle Filter [10] in a Markovian manner. Using stereo-based approach is possible to reduce the searching area and, consequentially, the elaboration time as described in [28, 25]. Examples of detections that are not based on images, but instead on time-of-flight (T.O.F.) like RADAR and LIDAR are put forth in [17, 38, 39]. These systems very often combine the T.O.F. sensor with a camera as Broggi et. al. [5] do with a combination of a NIR camera and a LIDAR. Furthermore they use a scenario-driven search approach where they only look for pedestrians in relevant areas. Further reading on pedestrian protection systems can be found in the survey by Gandhi and Trivedi [18], and comprehensive surveys on vision based detection systems are found in the papers [9, 12, 23, 19].

Table D.1: Key statistics about major public pedestrian datasets, courtesy of Dollár et. al [9]

	Training			Testing			Pedestrian height		
	# pedestrians	# negative images	# positive images	# pedestrians	# negative images	# positive images	10% percentile	median	90% percentile
MIT	924	-	-	-	-	-	128	128	128
INRIA	1208	1218	614	566	453	288	139	279	456
ETH	2388	-	499	12k	-	1804	50	90	189
TUD-Brussels	1776	218	1092	1498	-	508	40	66	112
Daimler-DB	15.6k	6.7k	-	56.5k	-	21.8k	21	47	84
Caltech	192k	61k	67k	155k	56k	65k	27	48	97

2.1 Public datasets

Several datasets are publicly available. The two best known are the MIT dataset [35] and the INRIA dataset [7], recently more comprehensive datasets have been put forth. These include the ETH [13], TUD-Brussels [42], Caltech [8], and Daimler Detection Benchmark (DaimlerDB) [12] pedestrian datasets. Note that the DaimlerDB set should not be confused with the older and smaller Daimler Classification Benchmark, often wrongly abbreviated DaimlerDB. Key stats about the datasets are presented in table D.1 as also presented by Dollár et al. [9]. While the INRIA dataset was used in the first presentation of this system [32], this paper deals mainly with the DaimlerDB, since that is a much larger dataset created with focus on in-car detection systems. All testing is done against the DaimlerDB (see section 4 for further details), and we compare trainings with the DaimlerDB and the INRIA dataset.

2.2 Performance of the state-of-the-art

In order to know what the performance target for a vision based system is, we turn to the evaluation of the state-of-the-art performance in [9]. Two results are interesting: the detection rate versus the false positive per frame and the detection speed (frame rate). As this paper uses the DaimlerDB pedestrian dataset, we compare our performance with the state-of-the-art detectors on this database, as reported in [9]. 10 different systems have been tested on the dataset and detection rates are available at a false positive rate of 0.1 false positives per frame (FPPF). The results can be seen in table D.2. Apart from the 10 systems which were tested on the DaimlerDB dataset, we have included the fastest detector of all. No detection results were reported for this detector on the DaimlerDB, but on other sets it achieved detection rates around 0.4.

Table D.2: Detection rates and speeds for state-of-the-art pedestrian detection systems at 0.5 FPPF on the DaimlerDB dataset, courtesy of Dollár et. al. [9]. The paper contains an explanation of each of the systems. These performances are directly comparable to the results obtained in this paper. The fastest system is also listed, although detection rates for the DC dataset are unknown. Abbreviations are the same described in [9]

Algorithm	Part-based	Detection rate	Speed
MultiFtr+Motion	no	0.75	0.004 FPS
LatSvm-V2	yes	0.69	0.164 FPS
MultiFtr+CSS	no	0.68	0.005 FPS
HogLbp	no	0.59	0.014 FPS
HikSvm	no	0.5	0.036 FPS
MultiFtr	no	0.49	0.017 FPS
LatSvm-V1	yes	0.48	0.098 FPS
HOG	no	0.42	0.054 FPS
Shapelet	no	0.10	0.010 FPS
VJ	no	0.09	0.089 FPS
FPDW	no	N/A for DC dataset	2.670 FPS

3 System overview

A two-stage system based on a combination of Haar-cascade classifier and a novel part-based HOG-SVM will be presented here; an innovative features-based pedestrian tracking approach will be also described.
A monocular vision system is used, since a simple on board camera is present in many new high-end cars already. A Haar detector is used to reduce the area of interest (ROI) - the *detection stage* - providing candidate pedestrians to the HOG detector which classifies the windows as pedestrians or non-pedestrians - the *verification stage*. To increase the robustness of the system and reduce the number of false positives, a part-based approach is used in the verification stage. The full body, upper body and lower body are each verified using a SVM. These three results are then combined to obtain the final response for the ROI. Two ways were investigated to combine results in the verification stage:

- a simple majority vote, where at least two of three SVMs must classify the window as pedestrian;
- a more advanced way, where another SVM classifies the window based on the estimated function value from a SVM regression performed on each part.

Due to the high variability in pedestrian appearance, a robust system with strict thresholds for detections may not detect the same pedestrian in subsequent frames and thus reduce the detection rate considerably. To counter this, a stage of feature-based tracking was introduced, significantly increasing the number of true positives.

Fig. D.1: The different bounding box required by Haar-cascade and HOG-SVM, base image from the DaimlerDB dataset [12]. The red and dashed line is Haar bounding box and the blue and continuous one is HOG bounding box

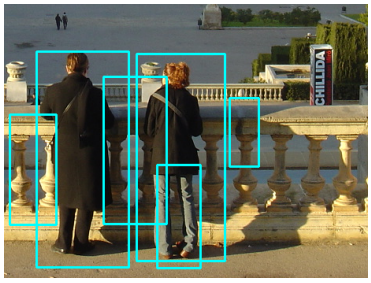


Fig. D.2: Detection stage output. Several false positives are contained but these will be removed in the verification stage.

3.1 Detection stage

An AdaBoost cascade on Haar-features is used in the detection stage. Several weak classifiers are combined into a strong one; the final classifier is formed with the combination of several layers of these strong classifiers. The cascade structure removes most false positive in the first stages, increasing the speed of the classifier not having to calculate these in following stages. Below we denote the number of cascade stages as k. [40] presents a comprehensive description of the algorithm. Unlike HOG features, Haar-like features do not benefit from having much background included. Training images need to be cropped closely around the annotated human shape, an example can be seen in Fig. D.1. Following the suggestions by Viola and Jones [40] about optimal image size for the Haar-cascade approach, the training images are resized to 20x40 pixels. Another interesting element in the training phase is the choice of data sets used to train the cascade classifier. Most of the older systems were trained with

Fig. D.3: Example of part boundaries for the 2- and 3-part verification.

the INRIA dataset, containing general environments and not specifically pedestrians. To show how the change in results with different training data sets, the system has been trained with the INRIA dataset alone, the Daimler-DB dataset alone, and also a combination of the two.

Since the detection stage defines the upper bound of detections for the entire system, it is fundamental to choose the best value for the number of the stages. A lower value of k means a high detection rate, but also a high number of false positives. Initially it might seem logical to choose the number of stages as low as possible, to ensure a high number of detections. That will, however, result in inaccurate bounding boxes (and many of them) as shown in D.4, and thus the final results will be incorrect. The part-based approach was not introduced to the detection stage, as preliminary tests, as well as the work by Alonso et al. [1] showed a bad performance for this approach. When the bounding boxes of candidate pedestrians (an example can be seen in Fig. D.2) have been obtained, they are passed to the verification stage.

3.2 Part verification stage

As opposed to the full-body verification stage of Geismann and Schneider's [22], a part-based detection scheme is used in this work. Two different compositions of body parts have been tested:

- full body, upper body and lower body;
- full body, head, torso and legs.

A fixed ratio between them have been used. Upper body and lower body are obtained dividing the shape into two equal parts. When we split the shape into three parts, instead, it was assumed a ratio of 16% for head and neck, 34% for torso, while legs are considered to occupy the 50% of the entire body. These numbers are taken from standard human body ratios. Before passing the ROIs to the SVMs, a preprocessing to add background and resize the image is needed to ensure good performance by HOG-SVMs, which takes some background into account. Then, the individual part verification and the combined verification forms the verification stage. A SVM regression based on dense HOG descriptors is calculated for each part in the ROIs given by the detection stage. Two different types of SVM were tested, a linear SVM and a non-linear SVM. Each was tested in two variants, binary or regression SVM.

Fig. D.4: Example of the degradation of the bounding box varying k from 13 in the last pictures to 9 in the second one and 8 in the first one.

The binary SVM provides only the classification (pedestrian or non-pedestrian) of the element; the last one provides the estimated function value. [22] uses a special kind of sparse HOG descriptors, whereas our algorithm uses classic dense HOG descriptors. Integral images were used to speed up the descriptor calculation as described in [37]. For SVM training, images from several datasets were tested with the goal of analyzing the effects of training sets in the verification stage. The process of training the SVMs for the different parts of the body are almost identical, the only changes being the portion of images used to calculate the HOG features. Examples of parts are shown in Fig. D.3.

3.3 Combined verification stage

For this last stage, two different approaches have been implemented: *majority vote* and *regression output classification*. The first one performing the final labeling without further classifiers; the last one uses one more classifier to label the window. There is a philosophical difference between the voting-based combination methods and the others: voting based combination require only a subset of body parts to be visible and detectable and can deal well with occlusions. The other requires all body parts to be visible, at least to some extent, so they will handle occlusions somewhat worse, but reduce the number of false positives. A possible compromise is to use occluded pedestrians in the dataset, training the classifier to detect pedestrians partially visible; obviously this also means an increase of false positive per frame.

The majority vote approach uses the binary outputs from the SVM. The value will be 1 if the classifier detects the specific part of the body or -1 if the part is not detected. A window is classified as correct detection if at least two out of three classifiers label the window as pedestrian. The formula used for the majority voting is:

$$l_{out} = \begin{cases} 1 & \text{if} \sum_{i=0}^{i<3} l_i >= 1 \\ -1 & \text{if} \sum_{i=0}^{i<3} l_i < 1 \end{cases} \tag{D.1}$$

where l_{out} is the final decision and l_i is the output from one of the three part-based detectors.

Regression output classification uses the three float value coming from SVMs of the verification stage to train a new classifier. Several type of classifiers were tested:

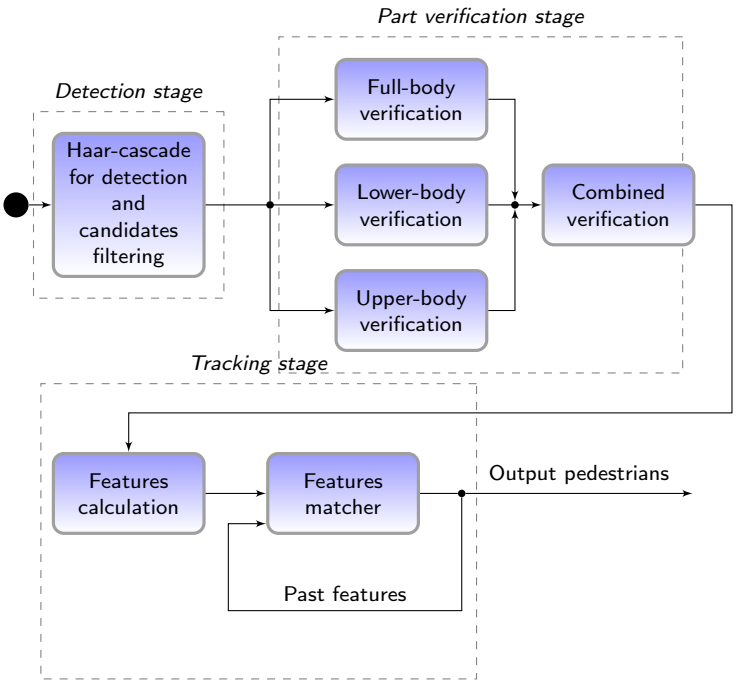


Fig. D.5: The flowchart of the algorithm described in this paper. Output of detection stage and following stages are bounding boxes.

a linear SVM, a non-linear SVM, and a Bayesian classifier; in the results section the different performances of each one will be shown.

3.4 Tracking stage

A feature-based tracking was used to enhance the detection rate. The tracker is introduced to increase the number of true positives due to the higher stability of the detection in case of for example occlusion and to decrease the number of false positives, since just the stable detection will be considered as pedestrians. The core of the tracking system is the feature matcher, using the matching approach of Geiger et al.[21]. The tracker labels pedestrians to supply possible missing detections due to mistakes of the classifier in the verification stage; a more detailed description of the tracking is presented in Section 5. An overview of the flow through the algorithm can be seen in Fig.D.5

4 Experiments

One of the main contributions of this paper is a thorough evaluation of the algorithm's parameters. This section describes experiments to determine the best detector, which

is then tested quantitatively and qualitatively in the next section. DaimlerDB was used primarily, with elements from the INRIA dataset in a few tests. Unless otherwise specified, images from the training part of DaimlerDB was used for training, both the detection stage and the part verification stage. The test part of DaimlerDB was split into two:

- one portion of 1500 images was used for the parameter optimization in this section;
- one portion of 500 images was used for the final test presented in the next section.

This ensures that the final performance measures are fully independent of the training images. The experiments are laid out like this:

1. The best detection stage training is determined and, then, the optimal value of k in the detection stage is decided.
2. The part-based verification is tackled with a comparison of the 2-part and 3-part approach. They are compared with a simple detector without part, similar to the original version of the algorithm as proposed by Geismann et. al. Furthermore the significance of each part is evaluated.
3. The combined verification stage is tested with various methods.
4. The system speed is tested and the time is broken down into individual stages.

4.1 PASCAL detection evaluation

For all the following experiments, the PASCAL measure [14] has been used to determine the detection rates. This is also used in [9], so the results should be directly comparable. The PASCAL measure evaluates to true if the overlap is more than 50%:

$$a_o \equiv \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{D.2}$$

where BB_{dt} and BB_{gt} are the bounding boxes of the detection and the bounding box of the ground truth, respectively. Each detection is compared with the ground truth of the 1500 images and counted as a true positive if a_o true, and as a false positive otherwise. All tests in the following are run on the complete system. For each test, all parameters are held fixed, except for the one in question. Thus, the results cannot necessarily be compared across tests, but the results are always comparable relative to each other within the tests.

4.2 Training of the detection stage

This test pitted different training setups of the Haar cascade. Four versions were tested:

- 2400 DaimlerDB images;
- 2400 INRIA images;
- 6000 images composed of 2400 INRIA and 3600 DaimlerDB images;

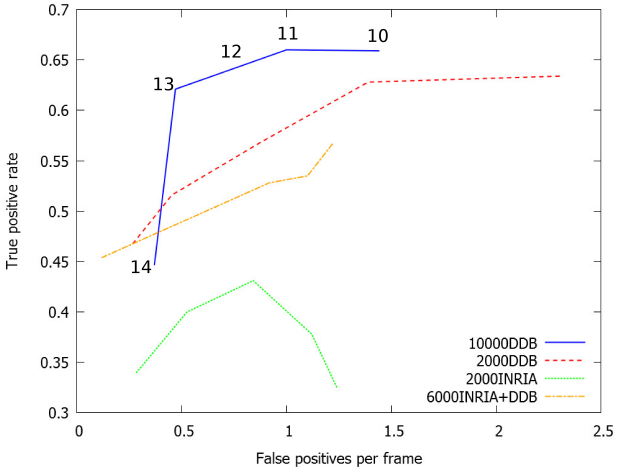


Fig. D.6: Comparison of different training sets for the detection stage. The system trained with DaimlerDB datasets performs significantly better remarking the excessive generality of INRIA. Chosen the best traning sets it was analyzed, for the system training with this dataset, the best value of k. As described in 4.3, 13 is the best one obtaining a good tradeoff between true positives and false positives.

- 10000 DaimlerDB images.

Results are presented in Fig. D.6 and show that performance is improved using more images. Fig. D.6 is a ROC curve created by plotting the fraction of true positives out of the positives (tpr = true positive rate) vs. the fraction of false positives out of the negatives (fpr = false positive rate), at various threshold settings. Note the bad performance of the system when trained with the INRIA datasets; this show how the INRIA are too general, being developed for the human detection. The big influence of this kind of dataset is also clearly visible in the system trained with 4000 DaimlerDB images and 2000 INRIA images; the system with less images (2000 DaimlerDB), but only from DaimlerDB dataset, performs better than this one with more images.

4.3 Choice of k in the detection stage

This test determines how many stages (k) the Haar cascade should have. As there are two verification stages after this, the detection stage should be tweaked so that it returns as many true positives as possible, whereas the number of false positives is less important; they will be removed later. Still, there is a point where raising the number of false positives does not provide a better detection performance, so the only effect will be a slow-down of the system since more ROIs must be inspected by the verification stages. Fig. D.6 shows ROC curve for different values of k. Few stages should mean raise the number of both false positives and true positives, but at some point the quality of the bounding boxes provided by the detection stages degrade to a level where the verification stage only verify a few candidates.

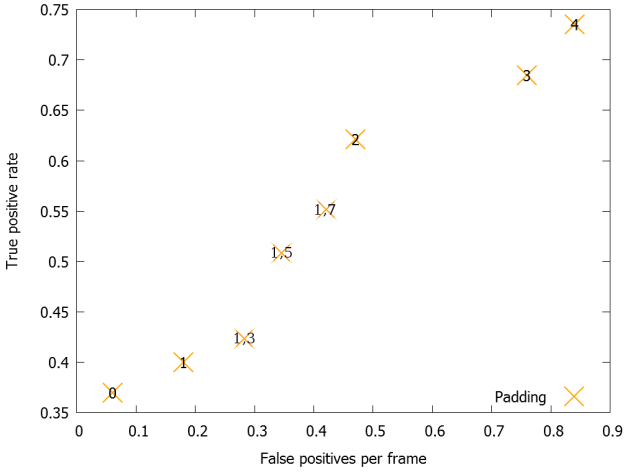


Fig. D.7: Choosing the padding put on the ROIs from the detection stage before passing them to the verification. Using the 10000 Daimler-DB training set and a k value of 13.

4.4 Part-verification padding

Padding p is the amount of area added to the ROIs returned by the detection stage. The HOG-SVM approach is sensitive to the amount of free space around the subject as described in [7], so the parameter is relevant for optimization. An example of padding is seen on Fig. D.2 where the bounding box for the Haar cascade is much closer to the subject than the rest. We express p as a fraction of the width of the ROI found by the detection stage:

$$p_{pixels} = \frac{w_{ROI}}{w_t} \cdot p \tag{D.3}$$

where p is the padding value, w_{ROI} is the width of the found ROI, w_t is the width of the training images, and p_{pixels} is the padding measured in pixels. Fig. D.7 shows the performance of different paddings. It is evident how less padding means worse images to the verification stage. At the same time, too much padding makes the verification more difficult for the HOG detector since more items are analyzed and more mistakes happen.

4.5 Number of parts

The performance of 1, 2, and 3-part verification is compared (with 1-part verification obviously not being part-based at all). Illustrations of the part boundaries for both the 2 and 3-part detector are shown on Fig. D.3. The performance of various part-numbers is shown on Fig. D.8.

2 parts is the best choice and the 3 parts performs better than 1 part at the lower false positive per frame. These results can be attributed to the quality of the images; the 3-part detector needs to detect the head, which is a comparatively small element and too hard to detect in a image with low resolution. With higher resolution images

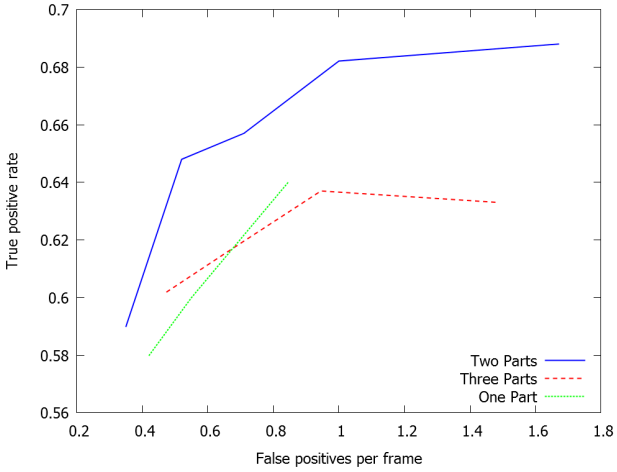


Fig. D.8: Detection performance with varying numbers of parts. Note that two parts out perform, while three parts are just as bad as one since the low quality of the images and the high difficulty in identifying small areas such as the head. Considering the results shown in the previous charts, has been used the 10000 Daimler-DB training set, a k value of 13 and a padding of 2.

it is likely that the 3-part approach would provide the best results, but at the same time the speed of the system would suffer.

In connection with this, an analysis of the significance of each part was done. The results show how the detection performance would be, relying on that specific part only. 4 parts have been tested: lower body, upper body, head, and torso. The lower body is used both for the 2- and 3-part verification, where the upper body is only used for the 2-part and the head and torso are used for the 3-part verification. Results of this analysis are shown on Fig. D.9. None of the parts alone perform better than a unified detector, but the upper body and torso provide the major contribution to the detection. These results support the hypothesis that the 3-part verification has a worse performance than the 2-part: due to the low resolution for the head detection. In this figure the head detection system is the worst, with a very low detection rate. The combination of upper body and legs/lower body is the best combination due to the high detection rate from the upper body and the reduction of false positives provided by the lower body.

4.6 Combined verification step

For the final combined verification step, four options have been investigated: linear SVM, radial SVM, and Bayesian classification for confidence classification and majority vote based on the discrete classification from the part-verifiers. The result of this comparison is shown in Fig. D.10. The vote-based combination should better deal with occlusions than the other approaches, but at the same time more false positives are returned by this method. The best performance - when the goal is a low false

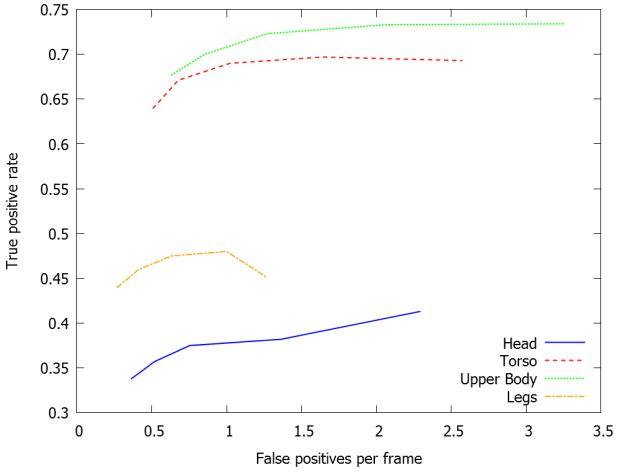


Fig. D.9: Detection performance with single parts, showing the reliability of each part type. The graph confirms the assumptions regarding the difficulty to detect the head. The same configuration parameters system of the last pictures it was used in this experiment.

positive per frame - is given by the radial approach. That follows logically from the non linearity of the data returned from the part detectors. The plot of the Bayesian approach shows excellent detection rate but with high number of false positives. Applying a linear separation on set of non-linear data, the Bayesian approach classifies more elements as pedestrians but, at the same time, incorrectly classifies a greater number of true negatives. This explains the high detection rate, but also the raise in false positives.

4.7 Speed evaluation

This test evaluates the speed of the system at various settings for the detection stage for a given hardware. The results are depicted in Fig.D.11. Changing k, the number of haar-cascade stages, has a large impact on the system speed, since it directly influences how many candidates the next stages must evaluate. The largest contribution in processing time is the full-body verification, while the contribution of the last stage is practically irrelevant. Setting a high k results in lower number of ROIs and a faster system, but also in a system capable of detecting fewer targets. The goal here is to choose the system where parameters are set to obtain a trade-off between speed and detection rate, taking the false positive per frame into account. Speed has been measured on a run of 1000 images and results are the mean of those. For the fastest one a complete calculation can be performed in about 0.757 s, corresponding to 1.32 frames per second.

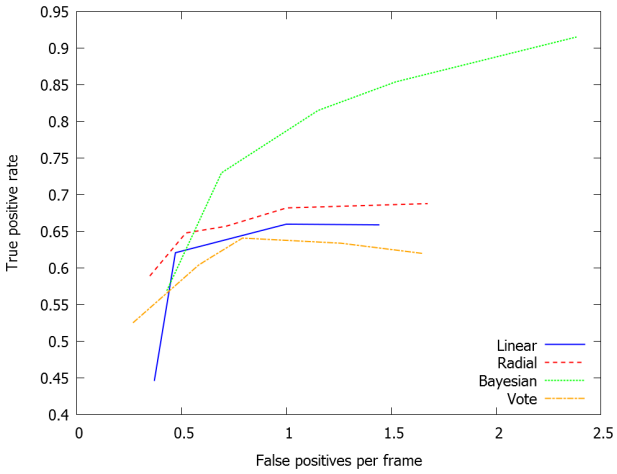
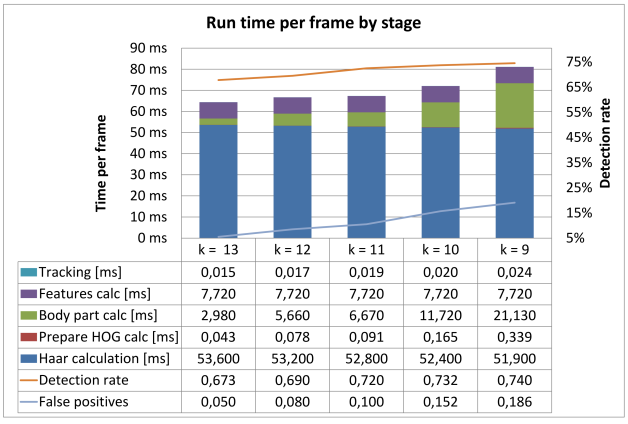


Fig. D.10: Comparison of different methods for combined part-verification. With a low false positive rate, the radial approach performs better. The system configuration is the following: Daimler-DB trainig set, k value of 13, padding value of 2 and two-part based approach.



	k = 13	k = 12	k = 11	k = 10	k = 9
Tracking [ms]	0,015	0,017	0,019	0,020	0,024
Features calc [ms]	7,720	7,720	7,720	7,720	7,720
Body part calc [ms]	2,980	5,660	6,670	11,720	21,130
Prepare HOG calc [ms]	0,043	0,078	0,091	0,165	0,339
Haar calculation [ms]	53,600	53,200	52,800	52,400	51,900
Detection rate	0,673	0,690	0,720	0,732	0,740
False positives	0,050	0,080	0,100	0,152	0,186

Fig. D.11: Speed versus detection rate and false positive per frame. The time has been measured for each stage, denoted as *k*. We see the reduction of false positives and the increase of true positive rising the number of stages. The PC is equipped with a Intel(R) Core(TM) i7-2670QM CPU @ 2.20GHz, and 8GB of DDR2 RAM.

4.8 Parametrization of the input image geometry

To make the system more configurable the possibility of choosing the images size has been added so processing time just can be adjusted by resizing the input image. Camera calibration parameters will automatically change to ensure the correct behavior of perspective and inverse perspective mapping functions used when filtering candidates. Resizing the image results in a reduction of processing speed and the true positives rate, as shown in table D.3.

4.9 Improvements over the old system

Significant improvements were applied to the system described in [32], as one can see in Fig.D.6 comparing the blue graph with the green one. At false positive per frame of 0.5 the TPR was increased from 0.4 to 0.63 with a speed-up of more than 16x. The filtering of candidates and feature-based tracking introduced a significant speed-up as the implementation was parallelized.

4.10 "Real-driving" experimental results

Figure D.12 shows some examples of possible circumstances that may occur in a real environment, varying from simple, medium and hard situations.
The first two lines represent simple situations with pedestrians crossing the street, ride bicycles or walking along the sidewalk and, some more critical situations, with pedestrians partially occluded in a structured environment. The line (C) shows, instead, some case of hard detection as highly occluded pedestrians, pedestrians underexposure and pedestrians situated in a highly complex scene.
A measure of the maximum distance of recognition is provided by the line (D), where you can see pedestrians recognized about 45/50 meters away.
Aside from it, our algorithm still has some shortcomings, as show in the last line. Some 'common' error of classification are shown in the last two pictures; however, these errors can be considered superficial, since only present in single frames and, therefore, detachable by our tracking system.
A relevant problem is shown in the first picture of line (E): due to the geometric filter on the size of pedestrians, our system does not detect pedestrians smaller than 1.45 m. A possible solutions is to broaden the constraints of the filter obtaining, then, a greater number of false positives. An additional downside of the geometric filter, regards the accuracy of the calculation of the inverse perspective mapping if the ground is not flat, as is presupposed by our system. Using techniques of image stabilization as described in [6, 44] could provide significant improvements; an alternative solution would involve the introduction of stereo based approach to filter candidates pedestrians.
A further case of interest is depicted in second illustration of the last line, illustrating the situation where a pedestrian is crossing the street where the car is turning. With cameras situated in the front of the vehicle, it is impossible to detect pedestrian in time to brake. A solution could be the introduction of cameras that allow to look on the side of the car and detect pedestrians in advance.

Fig. D.12: Line (A) shows examples of pedestrians detection with simple environment. Line (B), instead, shows examples of behavior of our detector in the presence of small occlusions and structured scenes. The potential of our classifier in the presence of pedestrians strongly occluded, highly structured scenes and underexposure of the camera is shown in line (C). Examples of detection of pedestrians far apart, about 45 m, are shown in line (D). Line (E) shows samples of possible detector problems: pedestrians too small, cyclist at the intersections and 'common' misclassification such as trees and poles.

5 Porting to a real prototype

With the aim of testing the developed system in the real world, the original stand alone software has been ported first to a prototyping software platform to optimize it in a laboratory setting, and then to a real hardware platform. Given on the results obtained on the real platform, a set of additional features have been identified and implemented to improve the detection performance in a number of critical situations. These modifications are described below.

5.1 Porting and optimization

The original code has been ported to be an application of the latest version of the GOLD[45] software.

GOLD offers a number of advantages in this phase, allowing the application to deal with virtual devices instead of using the hardware directly. This allows the system to work in the lab on recordings previously taken and stored on a disk, or on a real platform and taking data from the the hardware.

During the porting, a conversion of the Daimler database images has been done, making it possible to read this recording with GOLD and using this dataset as input for the pedestrian detection application. This has been done mainly for the availability of a high quality, per-frame ground truth that can be recovered any time and used to check the consistency of the results with those obtained with the standalone application.

GOLD also offers a profiling API allowing for timing of different parts of the application and see the time spent in the execution of these parts for every frame, as well as computing cumulative statistics collected across a playback session.

5.2 Platform

After reaching an acceptable performance level, the system was transferred to a real prototype vehicle [15].

The platform is equipped with 10 cameras, and one of these, looking forward, has been physically connected to the application. The camera used is a PointGrey DragonFly 2, working at 10Hz and producing images with a resolution of 1024x768. The camera is equipped with a 6mm micro lens, which provides an acceptable level of distortion for this application. The camera has a firewire interface which is connected to an adapter located in the trunk, in an industrial PC. The PC is equipped with a Intel(R) Core(TM) i7-2670QM CPU @ 2.20GHz, and 8GB of DDR2 RAM. Using this configuration and downsampling the image to 640×480 pixels it is possible to keep the processing time below 100ms, for simple scenes generating a reasonably low number of candidates. During the tests on the real platform some weaknesses of the original system emerged. These weaknesses were mostly due to a lacking of robustness of the results observed over long driving periods and include: a high number of false positives and two discontinuous recognitions of the same targets observed along several frames.

5.3 Candidates filtering

As a first step, a set of filters was introduced to reduce false positives to remove the candidates with size outside of a selected range $[1.45 - 2.20]$ m in real-world measurements. The IPM (Inverse Perspective mapping) technique [29, 4] was used to calculate the position of the pedestrian candidate in real world coordinates; by using the pedestrian baseline it is possible to determine the ratio pixel/meters at this distance and estimate the pedestrian height in the world knowing its height in image coordinates, using the flat road assumption. The application of this filter gives a good reduction of false positives with a small impact on a true positives; quantitative results are shown in the next section.

5.4 Features

Classification schemes can be enhanced with a tracking system to counteract the high instability of the detector due to the high variability of pedestrians. A feature-based tracking system was used to fix this lack: features provide a robust base to track people due to their translation and light invariance. A set of features, as detailed in 5.5, described in [21] based on multiple local convolution, key point and descriptors, are extracted from two different hash images. Stable feature locations are obtained filter the input images with 5x5 blob and corner masks and, then, it was applied non-maximum- and non-minimum-suppression [34] on the filtered images. Starting from pedestrians output from the verification stage, features are computed and used to match pedestrians in subsequent frames. The feature-based tracking has the downside of being dependent on the vehicle ego-motion. Vehicles moving at high speed, especially in conjunction with low frame rates, cause a high difference between two subsequent frames and, consequently, a bad match between corresponding features. To cope with this problem a higher frame rate must be used. Another downside of using features for a tracking system is the difficulty of distinguishing between foreground and background pixels. As a result, some matches could be wrong but the impact of these errors is very low and decrease pedestrian motion.

5.5 Tracking

When a candidate pedestrian has been recognized by the SVM for 250 ms (a time limit is used due to the variability of frame rates), it is considered a true pedestrian and it is introduced in the tracking system. In the following frames the pedestrian features will be matched with new candidate pedestrians and their positions and descriptor will be updated with the new one. It was compared a 11x11 block windows of horizontal and vertical Sobel filter responses to each other by using the Sum of Absolute Differences (SAD) error metric. The whole block window is reduce with Sobel responses to 8 bits and sum the differences over a sparse set of 16 locations. To further significant speed-up it was match only a subset of all features, found by Non-Maxima-Suppression (NMS). The feature are, then, assigned to a 50x50 pixel bin of an equally spaced grid and will be computed the minimum and maximum displacements for each bin. In this way we reduce the final search space and speed-up the system. If no candidate matches the search criteria (missing detection by the SVM), search for a match will

be done across the entire image. If a match is found, a ghost pedestrian will be introduced. It will be updated for up to 0.5 s, after which it will be removed. A flowchart of the tracking system can be seen in Fig. D.5.

5.6 Higher frame rate

The best results from the feature-based tracking are obtained in correspondence to a good match between the features extracted from the candidate images in consecutive frames. When working at low frame rates, like 10 Hz, the high variability between consecutive frames, both due to the object movement in the scene and the vehicle ego motion, leads to a bad performance of the feature matcher and, as a consequence, the tracking system. Using the prototype platform, a new set of images has been recorded at 30 Hz from one of the forward looking cameras. These images have been used offline with tracking enabled showing significant improvements in the result robustness, reducing the blinking of correct detections caused by missed detections in single frames, and also the false positives. Unfortunately, the framerate of the the Daimler-DB dataset is lower than 10 Hz and this is a limiting factor for comparing recognition performance improvement with the tracking system. To get a significant sampling speed in real time, the prototype was altered to acquire images with different size. The reduction of the input image to 320×240 pixels leads to a framerate of 20 fps, and offers a level of recognition performance similar to the one obtained at 30 Hz with the offline processing.

6 Final performance evaluation

After the evaluation of all the parameters a final system to be tested on the DaimlerDB has been defined. Table D.4 contains the parameters values used in the final system.

6.1 Final test without tracking

Figure D.13 shows the performance of the final system. This figure shows results for several values of k in order to plot ROC curves, and gives a detection rate of about 0.69 with a false positive per frame of 0.5 considering 13 as the best value for k. Despite a high false positive per frame, our system is directly comparable with others shown in [9]; it shows the same performance of *LatSvm-V2*, one of the most successful part-based approach described in [16], but with a huge speedup of 10x (not considering the extra speedup described below). Better performance is achieved by filtering the candidates as described in the previous section, reducing the false positive rate from 0.5 to 0.046 with a small reduction of true positive rate to 0.673 as shown in Fig.D.14. These results allow our system to gain a foothold in the state of the arts consolidated by a huge speedup described below. The results are summarized in table D.3.

6.2 Tracking improvements

Introducing a tracking system resulted in significant improvements in the number of true positives and a reduction of the number of false positives. The performance

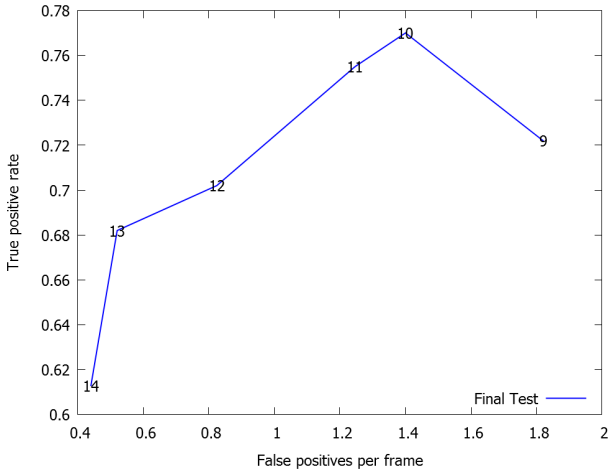


Fig. D.13: Final test detection. Evaluation on performance on the last part of Daimler-DB images without the optimization for the porting on a real prototype.

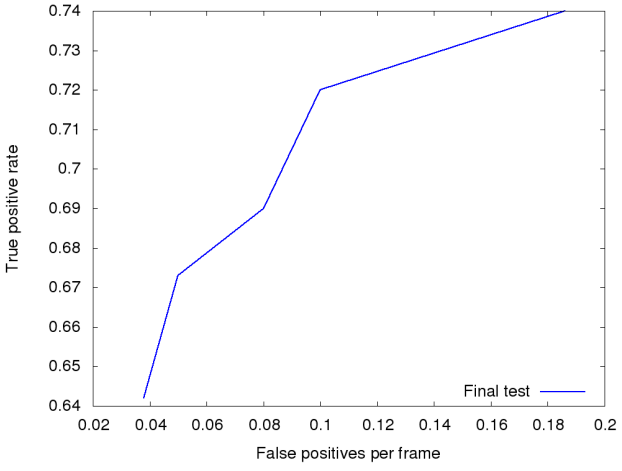


Fig. D.14: Evaluation on performance of the optimized system on the last part of Daimler-DB images.

Table D.3: Final system performance

	Detection rate	False positive per frame
Basic system	0.69	0.5
With candidate size filtering	**0.673**	**0.046**
With resized image (320x240)	0.55	0.02

	Frame rate	Image size
Unoptimized system	1.32 FPS	640x480
Parallelized system	**16.67 FPS**	640x480
	30 FPS	320x240
Fastest system in [9]	2.6 FPS	640x480

Table D.4: Final system configuration

Training Dataset	10000 Daimler-DB
k-value	13
Padding	2
Number of parts	2
Combined verification	Radial

improvements due to the introduction of tracking were tested on our own dataset (two sequences of 5182 and 11490 frames respectively) captured on the real prototype described in section 5. It was not possible to use the DaimlerDB test set due its low frame rate of about 10 Hz, too low to ensure a stable tracking. An increase of 27% and 22% of true positives on the two dataset was obtained with a reduction of 5% and 10% of false positives. These results showcase the better stability of the system, allowing to track the pedestrian in consecutive frames and opening the way to further improvements such as determining pedestrian direction and orientation [20].

6.3 Performance on the prototype platform

In order to guarantee real-time performance on the prototype platform (GOLD), a parallelization techniques was introduced. Parallelization of Haar-features and HOG-features calculation and classification were obtained by compiling OpenCV with TBB (Thread Building Blocks) enabled. In this way it is possible to take advantage of multicore CPUs. A further parallelization was obtained by executing the classification of HOG features for the different body parts on separate threads, reducing the verification stage processing time of about 30%. With an image of 640x480 pixels the processing time changed from 755 ms to 60 ms, a 12x speed-up. Thus, our system is running 8 times faster than the fastest system presented in [9]; 16.67 fps versus 2.6 fps. A further speed-up can be provided by reducing the images size, which results in a

Fig. D.15: Examples of detection on a prototype platform. People in different poses are detected, including cyclists or people walking close a tree often hard to detect. A missed detection is shown in the last figure due to overexposure of the pedestrian.

processing speed of about 30Hz on an image of 320x240. This approach, however, has a detrimental effect on detection rates. Examples of detection on a prototype platform can be seen in Fig. D.15

7 Future work

Various studies are currently ongoing in order to improve the presented pedestrian detector. Since feature based tracking works better at higher frame rates, a low level reimplementation of the two stage classifier fully exploiting multicore processors (or GPU) features may give some significant speed up. The current system relies on OpenCV 2.4 compiled with Intel Thread Building Blocks support. Looking at the CPU utilization, we get values between 60-80% for each core, which is a clear indication that some serial piece of code is still present. Reducing the image area, the processor utilization falls, ranging from 80% at 640×480 pixels to 60% for 320×240 pixel images.

Another improvement can be added to the high level processing, introducing filters on the predicted pedestrian trajectory. Especially when working with high frame rates, a good tracking of the pedestrian trajectory is produced from the current system. A Kalman filter could provide a prediction of the trajectory that pedestrian is taking in the future, which could be evaluated to predict dangerous situations.

The vehicle ego motion has intentionally not used for this system, since one of the constraints was to obtain a final system simply relying on vision. Introducing a visual odometry block could supply information on ego motion without breaking this requirement. However additional computational power would be needed.

8 Concluding remarks

In this paper a novel pedestrian detector system, running on a prototype vehicle platform has been presented. The algorithm generates possible pedestrian candidates from the input image using an Haar cascade classifier. Candidates are then validated

through a novel part based HOG filter. A feature based tracking system takes the output of the two-stage detector and compares the features of new candidates with those of the past. A matching is performed with the aim of assigning a consistent label to each candidate and improving the recognition robustness, by filling false negative filtered by the previous phases. The whole system has been ported to a prototyping framework and integrated on a platform vehicle, for testing and optimizations. A significant performance improvement has been obtained exploiting the CPU multicore features. As result a system working at 20Hz and offering performance comparable to the state-of-the-art has been obtained. Additional real world tests have been performed on the platform for finding weaknesses. Even though the system is fast compared to the state-of-the-art, its detection performance compares very favorably to the state-of-the-art with a true positive rate of 0.673 at a false positive per frame of only 0.046.

Acknowledgments

The authors would like to thank our colleagues in the LISA-CVRR lab. The authors would also like to acknowledgment Cassa di Risparmio di Parma e Piacenza for funding the test platform used for this work.

Bibliography

[1] I.P. Alonso, D.F. Llorca, M.A. Sotelo, L.M. Bergasa, P.R. de Toro, J. Nuevo, M. Ocaña, and M.A.G. Garrido. "Combination of feature extraction methods for SVM pedestrian detection". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.2 (2007), pp. 292–307.

[2] M. Andriluka. "People-tracking-by-detection and people-detection-by-tracking". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (June 2008), pp. 1–8.

[3] S. Avidan. "Ensemble tracking". In: *PAMI* (2007), pp. 261–271.

[4] Massimo Bertozzi, Alberto Broggi, Alessandra Fascioli, and Ra Fascioli. "Stereo Inverse Perspective Mapping: Theory and Applications". In: *Image and Vision Computing Journal* 8 (1998), pp. 585–590.

[5] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H.G. Jung. "A new approach to urban pedestrian detection for automatic braking". In: *Intelligent Transportation Systems, IEEE Transactions on* 10.4 (2009), pp. 594–605.

[6] A. Broggi, P. Grisleri, T. Graf, and M. Meinecke. "A software video stabilization system for automotive oriented applications". In: *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st* 5 (May 2005), pp. 2760–2764. ISSN: 1550-2252. DOI: 10.1109/VETECS.2005.1543849.

[7] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*. 2005.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona. "Pedestrian detection: A benchmark". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. June 2009, pp. 304–311. DOI: 10.1109/CVPR.2009.5206631.

[9] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. "Pedestrian Detection: An Evaluation of the State of the Art". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.4 (2012), pp. 743–761.

[10] Arnaud Doucet and Freitas. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001. ISBN: 978-0-387-95146-1.

[11] M. Enzweiler and D.M. Gavrila. "Integrated pedestrian classification and orientation estimation". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. June 2010, pp. 982–989. DOI: 10.1109/CVPR.2010.5540110.

[12] M. Enzweiler and D.M. Gavrila. "Monocular pedestrian detection: Survey and experiments". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.12 (2009), pp. 2179–2195.

[13] A. Ess, B. Leibe, and L. Van Gool. "Depth and Appearance for Mobile Scene Analysis". In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. Oct. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4409092.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.

[15] Paolo Grisleri Fedriga and Isabella. "The BRAiVE platform". In: *Procs. 7th IFAC Symposium on Intelligent Autonomous Vehicles*. Lecce, Italy, Sept. 2010.

[16] P. Felzenszwalb, D. McAllester, and D. Ramanan. "A discriminatively trained, multiscale, deformable part model". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (June 2008), pp. 1–8. ISSN: 1063-6919. DOI: 10.1109/CVPR.2008.4587597.

[17] K.C. Fuerstenberg. "Pedestrian protection using laserscanners". In: *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*. Sept. 2005, pp. 437–442.

[18] T. Gandhi and Mohan M. Trivedi. "Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation". In: *Machine Vision Applications* 18.3-4 (2007), pp. 207–220.

[19] Tarak Gandhi and Mohan M. Trivedi. "Computer Vision and Machine Learning for Enhancing Pedestrian Safety". In: *Computational Intelligence in Automotive Applications*. Ed. by Danil Prokhorov. Vol. 132. Studies in Computational Intelligence. Springer Berlin / Heidelberg, 2008, pp. 79–102. ISBN: 978-3-540-79256-7. URL: http://dx.doi.org/10.1007/978-3-540-79257-4_4.

[20] Tarak Gandhi and Mohan M. Trivedi. "Image based estimation of pedestrian orientation for improving path prediction". In: *Intelligent Vehicles Symposium, 2008 IEEE* (June 2008), pp. 506–511.

[21] Andreas Geiger, Julius Ziegler, and Christoph Stiller. "Stereoscan: Dense 3d reconstruction in real-time". In: *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE. 2011, pp. 963–968.

[22] P. Geismann and G. Schneider. "A two-staged approach to vision-based pedestrian recognition using Haar and HOG features". In: *Intelligent Vehicles Symposium, 2008 IEEE*. IEEE. 2008, pp. 554–559.

[23] D. Geronimo, A.M. Lopez, A.D. Sappa, and T. Graf. "Survey of pedestrian detection for advanced driver assistance systems". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.7 (2010), pp. 1239–1258.

[24] David Gerónimo, Angel D. Sappa, Antonio López, and Daniel Ponsa. "Adaptive Image Sampling and Windows Classification for On-board Pedestrian Detection". In: *ICCV*. University of Bielefeld, 2007.

[25] W. Khan and J. Morris. "Safety of stereo driver assistance systems". In: *Intelligent Vehicles Symposium (IV), 2012 IEEE* (June 2012), pp. 469–475.

[26] S.J. Krotosky and Mohan M. Trivedi. "On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.4 (Dec. 2007), pp. 619–629. ISSN: 1524-9050. DOI: 10.1109/TITS.2007.908722.

[27] B. Leibe, K. Schindler, and L. Van Gool. "Coupled Detection and Trajectory Estimation for Multi-Object Tracking". In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (Oct. 2007), pp. 1–7.

[28] D. F. Llorca, M.A. Sotelo, A.M. Hellín, A. Orellana, M. Gavilán, I.G. Daza, and A.G. Lorente. "Stereo regions-of-interest selection for pedestrian protection: A survey". In: *Transportation Research Part C: Emerging Technologies* 25 (Dec. 2012), pp. 226–237.

[29] H A Mallot, H H Bülthoff, J J Little, and S Bohrer. "Inverse perspective mapping simplifies optical flow computation and obstacle detection." In: *Biol Cybern* 64.3 (1991), pp. 177–85. ISSN: 0340-1200. URL: http://www.biomedsearch.com/nih/Inverse-perspective-mapping-simplifies-optical/2004128.html.

[30] X. Mao, F. Qi, and W. Zhu. "Multiple-part based Pedestrian Detection using Interfering Object Detection". In: *Natural Computation, 2007. ICNC 2007. Third International Conference on*. Vol. 2. IEEE. 2007, pp. 165–169.

[31] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. "Human Detection Based on a Probabilistic Assembly of Robust Part Detectors". In: *Computer Vision - ECCV 2004*. Ed. by Tomás Pajdla and Jirí Matas. Vol. 3021. Lecture Notes in Computer Science. 10.1007/978-3-540-24670-1_6. Springer Berlin / Heidelberg, 2004, pp. 69–82. ISBN: 978-3-540-21984-2.

[32] Andreas Møgelmose, Antonio Prioletti, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund. "Two-Stage Part-Based Pedestrian Detection". In: *15th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sept. 2012, pp. 73–77.

[33] A. Mohan, C. Papageorgiou, and T. Poggio. "Example-based object detection in images by components". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (Apr. 2001), pp. 349–361.

[34] A. Neubeck and L. Van Gool. "Efficient Non-Maximum Suppression". In: *18th International Conference on Pattern Recognition (ICPR)* (2006), pp. 850–855.

[35] Constantine Papageorgiou and Tomaso Poggio. "A trainable system for object detection". In: *International Journal of Computer Vision* 38 (1 2000), pp. 15–33. ISSN: 0920-5691.

[36] D. Park, D. Ramanan, and C. Fowlkes. "Multiresolution models for object detection"". In: *European Conference Computer Vision* (2010).

[37] F. Porikli. "Integral histogram: A fast way to extract histograms in cartesian spaces". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 829–836.

[38] U. Scheunert, H. Cramer, B. Fardi, and G. Wanielik. "Multi sensor based tracking of pedestrians: a survey of suitable movement models". In: *Intelligent Vehicles Symposium, 2004 IEEE*. June 2004, pp. 774–778. DOI: 10.1109/IVS.2004.1336482.

[39] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata. "Pedestrian detection with convolutional neural networks". In: *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*. June 2005, pp. 224–229.

[40] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* 1 (2001), pp. 511–518. ISSN: 1063-6919.

[41] Paul Viola, Michael J. Jones, and Daniel Snow. "Detecting Pedestrians Using Patterns of Motion and Appearance". In: *International Journal of Computer Vision* 63 (2 2005), pp. 153–161. ISSN: 0920-5691. URL: http://dx.doi.org/10.1007/s11263-005-6644-8.

[42] C. Wojek, S. Walk, and B. Schiele. "Multi-cue onboard pedestrian detection". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. June 2009, pp. 794–801. DOI: 10.1109/CVPR.2009.5206638.

[43] B. Wu and R. Nevatia. "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors". In: *International Journal of Computer Vision* 75.2 (2007), pp. 247–266.

[44] Luca Bombini Zani, Pietro Cerri, Paolo Grisleri, Simone Scaffardi, and Paolo. "An Evaluation of Monocular Image Stabilization Algorithms for Automotive Applications". In: *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2006* (Sept. 2006), pp. 1562–1567. eprint: doi:10.1109/ITSC.2006.1707446.

[45] Massimo Bertozzi Zani, Luca Bombini, Alberto Broggi, Pietro Cerri, Paolo Grisleri, and Paolo. "GOLD: A framework for developing intelligent-vehicle vision applications". In: *IEEE Intelligent Systems* 23.1 (Jan. 2008), pp. 69–71. eprint: doi:10.1109/MIS.2008.6.

[46] Li Zhang, Bo Wu, and R. Nevatia. "Pedestrian Detection in Infrared Images based on Local Shape Features". In: *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*. June 2007, pp. 1–8.

[47] L. Zhe, H. Gang, and L. Davis. "Multiple instance fFeature for robust part-based object detection". In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (June 2009), pp. 405–412.

Paper E

Attention Estimation by Simultaneous Analysis of Viewer and View

Ashish Tawari, Andreas Møgelmose, Sujitha Martin, Thomas B. Moeslund, and Mohan M. Trivedi

The paper has been published in the
Proceedings of the 17th International Conference on Intelligent Transportation Systems (ITSC), pp. 1381–1387, 2014.

The layout has been revised.

Abstract

This paper introduces a system for estimating the attention of a driver wearing a first person view camera using salient objects to improve gaze estimation. A challenging data set of pedestrians crossing intersections has been captured using Google Glass worn by a driver. A challenge unique to first person view from cars is that the interior of the car can take up a large part of the image. The proposed system automatically filters out the dashboard of the car, along with other parts of the instrumentation. The remaining area is used as a region of interest for a pedestrian detector. Two cameras looking at the driver are used to determine the direction of the driver's gaze, by examining the eye corners and the center of the iris. This coarse gaze estimation is then linked to the detected pedestrians to determine which pedestrian the driver is focused on at any given time.

1 Introduction

First person or ego-centric vision attempts to understand human behavior by acquiring information on what the person is looking at [11]. It employs videos/images from head mounted cameras. Recently, technological advances have made lightweight, wearable, egocentric cameras both practical and popular in various fields. The GoPro camera for instance can be mounted on helmets and is popular in a lot of sports such as biking, surfing, and skiing. The Microsoft SenseCam can be worn around the neck and has enough video storage to capture an entire day for the idea of "life logging". Cognitive scientists like to use first-person cameras attached to glasses (often in combination with eye trackers such as Tobii or SMI) to study visual attention in naturalistic environments. Most recently, emerging products like Google Glass have begun to make the first attempts to bring the idea of wearable, egocentric cameras into the mainstream.

Advances in the wearable-devices have enabled novel data acquisition in real-world scenarios. In the field of egocentric video, much of the recent work has focused on object detection, first-person action and activity detection, and data summary in context of "life-logging" video data. In this work, we present a unique data set collected during complex driving tasks with the aim of understanding driver-state and driver-'attention'. We use Google Glass to capture the driver's field of view, and a distributed camera setup instrumented in the vehicle to observe the driver's head and eye movements. Wearable and uncluttered cameras provide a practical advantage of ease of capture. The challenge, however, in our distributed camera setup, is to acquire data in sync to understand the driver-state, the environment and the vehicle-state simultaneously.

Driver gaze and head-pose are linked to the driver's current focus of attention [14, 6]. Therefore, eye and/or head tracking technology has been used extensively for visual distraction detection. The driving environment presents challenging conditions for a remote eye tracking technology to robustly and accurately estimate eye gaze. Even though precise eye gaze is desirable, coarse gaze direction is often sufficient in many applications [21]. By having a head mounted camera, we have direct access to the field of view of the driver. By analyzing salient regions in the field of view

(bottom-up attention model), one can estimate the focus of attention of the driver [2]. However, in a complex task such as driving, it is hard to say precisely where or at what we are looking, since eye fixations are often governed by goal-driven mechanisms (top-down attention model).

Towards this end, we propose a Looking-In-Looking-Out framework to estimate the driver's focus of attention by simultaneously observing the driver and the driver's field of view. We propose to measure coarse eye position and combine the salience of the scene to understand what object the driver is focused on at any given moment. We are not proposing a precise gaze tracking approach, but rather to determine the driver's attention by understanding coarse gaze direction and combining it with analysis of scene salience to determine important areas of interest - in our case pedestrians.

Our interest in pedestrians comes from the fact that in 2011, pedestrian deaths accounted for 14 percent of all traffic fatalities in motor vehicle traffic crashes in the United States. Almost three-fourths (73%) of pedestrian fatalities occurred in an urban setting versus a rural setting. 88% of pedestrian fatalities occurred during normal weather conditions (clear/cloudy), compared to rain, snow and foggy conditions. By knowing which pedestrians the driver has and has not seen, measures against collisions can be taken more accurately. While our main focus is on pedestrians, the framework can easily accommodate any object of interest or even a low-level saliency model to estimate the focus of attention.

The remainder of the paper is organized as follows. We give an overview of relevant related work in section 2 and explain the methods for determining gaze and detecting pedestrians in section 3. In section 4 we briefly review our captured data, and section 5 shows our results, before wrapping up with some concluding remarks and future work in section 6.

2 Related Work

Use of wearable cameras is not new [18]. In the last decade, gaze tracking systems such as [5], Tobii and SMI have made mobile gaze tracking in real life settings possible. More recently, the advances in hardware technology have made their usage more common in the computer vision community [19, 12, 17, 8, 13]. These systems are often used successfully in laboratory or controlled lighting conditions. Their use in complex environments is limited due to lengthy calibration, motion, illumination changes and, in case of driving, possible hindrance to the driver's front- or side-view. We discuss select work in activity recognition and gaze-behavior related research areas which are relevant in our current and larger interest in studying driver intent and behavior in real-world driving.

Ogaki et al. [15], using an inside-out camera system, combined eye-motion from inside looking camera and global motion from outside one to recognize indoor office activities. The authors suggest that joint cues from inside looking and outside looking cameras perform the best across different users. Doshi and Trivedi [6] introduced a similar system, but primarily for vehicular use. Pirsiavash and Ramanan [16] detected indoor apartment activities of daily living in first person camera view. They used object-centric action models which perform much better than low-level interest points based one to recognize activities. They show that using ground-truth object labels

in the action models significantly improves recognition performance. This suggests that recognizing objects of interest is key to recognizing tasks/activities in naturalistic settings.

Gaze allocation models are usually derived from static picture viewing studies. Many of the existing works are based on the computation of image salience [9] using low-level image features such as color contrast or motion to provide a good explanation of how humans orient their attention. However, these models fail for many aspects of picture viewing and natural task performance. Borji et al. [1] observe that object-level information can better predict fixation locations than low-level saliency models. Judd et al. [10] show that incorporating top-down image semantics such as faces and cars improves saliency estimation in images.

Inspired by the above findings, we present a driver's visual attention model using inside and outside looking camera views. In particular, we propose a model to determine coarse gaze direction and combine it with an object based saliency map to determine the allocated attention of the driver. Note that our interest lies in 'higher-level' semantic information about the driver attention and not 'low-level' precise gaze measurement. Our proposed framework circumvents the precise gaze estimation problem by utilizing a saliency map to achieve robust performance. Precise eye gaze from remote cameras is difficult not only due to low resolution of the eye region, but also due large head turns, self occlusion, illumination changes and hard shadows existing in an ever changing dynamic driving environment. To deal with large head turns and self occlusion, we propose to use a distributed camera system to monitor the driver.

We evaluate the proposed framework using a novel naturalistic driving data set using multiple cameras monitoring the driver and the outside environment. We use the head mounted camera from Google Glass to capture the driver's field of view. This particular device did not provide the ability to automatically synchronize footage with other cameras at per frame level. However, the ease, quick setup time (wearing and pressing capture button) as well as clean and uncluttered face view still makes the device a good choice. To obtain frame level synchronization, we mount an outside looking camera on the ego-vehicle which in turn is synchronized to the rest of the systems. Details on our synchronization strategy is provided in section 4.A head mounted camera provides the ability to capture not only the driver's outside field of view but also inside cockpit-view. In this work, we focus on the analysis of the outside view using the head mounted camera. This view poses unique challenges as discussed later.

3 Attention Estimation: LILO Framework

To infer the driver's attention, we are interested in knowing what object, in our case which pedestrian, the driver is looking at. There are two steps involved: first, estimating where driver is looking and second, detecting objects of interest in his/her field of view.In our current analysis, we focus on horizontal gaze variation since that is the most volatile and exercised direction by the driver to gain the knowledge of the environment. As we motivated earlier, we only require coarse gaze-direction and to distinguish it from precise gaze-value, we call it gaze-surrogate.

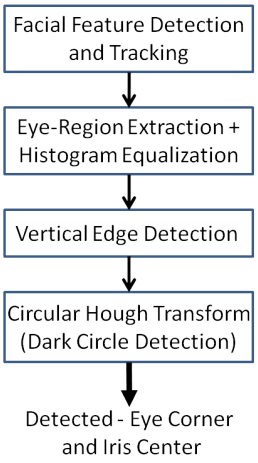


Fig. E.1: Block diagram for extracting iris center

3.1 Gaze-Surrogate Estimation

We automatically detect facial features - eye corners and iris center, and use cylindrical eye-model E.2 to estimate coarse gaze-direction. We use a facial feature tracking approach similar to [22] for detection of eye corners. During driving, however, large out-of-plane rotation of the head severely degrades the tracking performance. Hence, we use a two camera-system as proposed by Tawari et al. [20] to continuously track the facial features. From the facial features, we also calculate head pose, to be used in the gaze-direction calculation as explained below. We encourage the reader to refer to [22] and [20] for the details about eye-corner tracking, and head pose estimation and camera hand-off procedures. Here, we detail the iris detection and gaze-direction estimation algorithms.

Iris detection: The most prominent and reliable features within the eye region are the edges of the iris. The upper and lower eyelids in real face images occlude parts of the iris contours. Only the unoccluded iris edges can be used to fit the iris contour in the image plane. We detect the iris edge using a vertical edge operator in between upper and lower eyelids. The iris contours on the image plane are simplified as circles and center of the iris is detected using the circular Hough transform. Figure E.1 shows the block diagram of the iris detection algorithm.

Gaze-direction: Once the iris center is detected in the image plane, the gaze-direction β with respect to the head, see figure E.2, is estimated as a function of α, the angle subtended by an eye in the horizontal direction, head-pose (yaw) angle θ, and the ratio of the distances of iris center from the detected corner of the eyes in the image plane. Equation E.1-E.2 shows the calculation steps.

$$\frac{d_1}{d_2} = \frac{\cos(\theta - \alpha/2) - \cos(\theta - \beta)}{\cos(\theta - \beta) + \cos(180 - \theta - \alpha/2)} \tag{E.1}$$

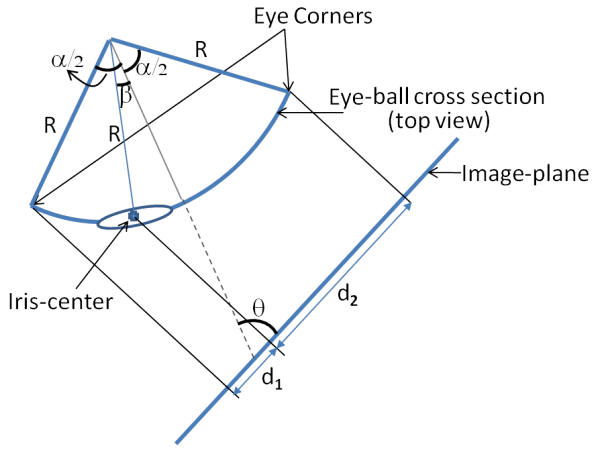


Fig. E.2: Eye ball image formulation: estimating β, gaze-angle with respect to head, from $alpha$, θ, d_1 and d_2

$$\beta \quad = \quad \theta \; - \; \arccos\left(\frac{2}{d_1/d_2+1}\sin(\theta)\sin(\alpha/2) \; + \; \cos(\alpha/2 \; + \; \theta)\right) \qquad \text{(E.2)}$$

Since the raw eye-tracking data is noisy (due to blinking and tracking errors), we smooth angle β with a median filter.

3.2 Salient Object Detection

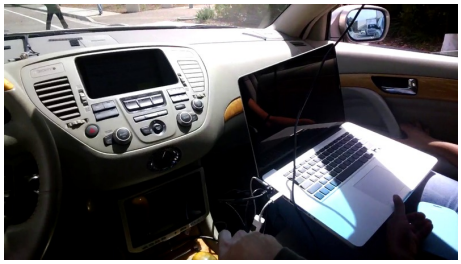
focus of attention detection in this paper is on pedestrians, thus requiring a pedestrian detector. Using first person view presents a number of interesting challenges compared to a stationary car-mounted camera. The major challenge is to determine the region of interest in which to look for pedestrians. With a stationary camera, it is either mounted so there are no obstructions in its view of the road, or it is mounted so any obstructions can easily be masked out manually. This is not the case for first person view, where the perspective constantly changes and there is no way of setting up a constant mask. This section introduces an algorithm to automatically mask out the dashboard and other unwanted areas.

The pedestrian detection module in this system is based on the classic HOG-SVM detection presented by Dalal and Triggs in [4]. It is trained on the Inria person dataset from the same paper. The pedestrian detection itself is simply a module in the full system, and it could be swapped with other approaches without issues.

The most important part of the interior mask is the dashboard mask. The dashboard can take up just the bottom of the image, the majority of the image, or not be present at all (fig. E.3) and the algorithm must handle all of those situations. We detect the distinct line between the windshield and dashboard and build from that:

1. Smooth out the input image with a Gaussian blur to even out noise.

(a) (b)



(c)

Fig. E.3: Three dashboard images showing examples of the very unconstrained position and orientation the dashboard can have in the field of view.

2. Detect edges using the Canny edge detector [3].
3. Determine the major lines in the image using the generalized Hough transform [7].
4. Filter the lines by angle to include only near-horizontal lines.
5. Build a confidence map of the dashboard.

Fig. E.4 shows sample output of step 4. Green lines are those that are horizontal enough to be considered in the dashboard map, red lines are ignored due to their extreme angles.

For each detected line, a polygon is drawn, which masks out all of the image below the line. These masks are combined and result in a single-frame dashboard map. To counter noisy line detections, a cumulative confidence map is introduced.

The cumulative confidence map is created by adding 1 to all pixels in the map covered by the current single-frame map and subtracting 1 from all pixels not covered by the current single-frame map. Areas that are detected in several subsequent frames will grow to a high confidence, but after a while of no detections, the confidence will fall and eventually the mask disappears. Examples of confidence maps and masks are in fig. E.5.

The use of the cumulative map is governed by two parameters, κ and λ. κ is the mask threshold. Any pixel in the confidence map with a value higher than κ is considered part of the dashboard map. In this implementation $\kappa = 2$. This parameter

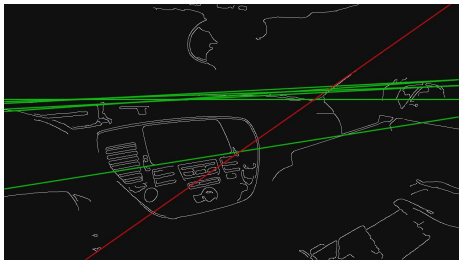


Fig. E.4: Detected lines in the image. Red lines are discarded due to too much of a skew to constitute the dashboard edge.

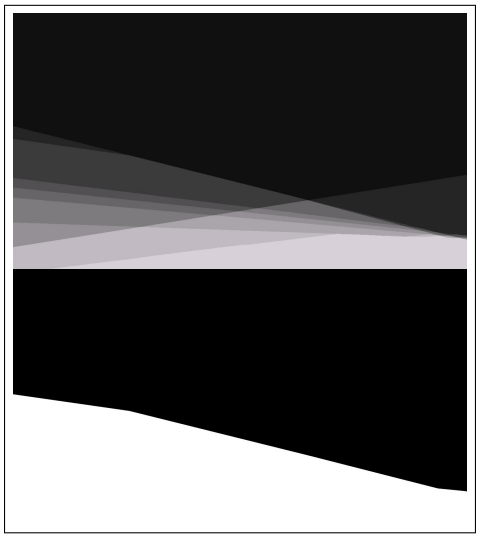


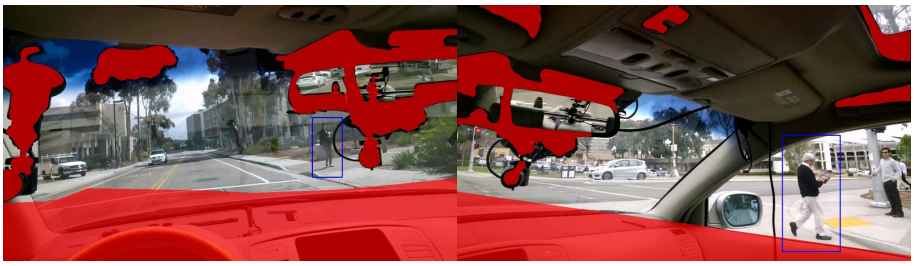
Fig. E.5: Confidence map (top) and its resulting mask (bottom).

controls how confident the system must be in a given pixel to include it in the mask. λ is the upper limit of confidence values.For a very high λ value, the confidence can grow very high, thus resulting in a long delay before the pixel goes below $\kappa.\lambda$ defines how long the memory of the system is. In this implementation $\lambda = 10$.

Apart from filtering out the dashboard, we detect and filter out black blobs large enough that they can only be part of the interior. We also discard pedestrian bounding boxes larger than 40% of the frame height.

3.3 Attended Object Determination

This step combines gaze-direction (β) and the salient object detected to determine which object the driver is attended to. This requires mapping from gaze-direction with respect to the head, β to pixel position in the external looking camera image. Equation E.3 shows the mapping function to determine the x-position (i.e. the yaw-



(a) (b)

Fig. E.6: Examples of pedestrian detection scenarios where the exclusion mask has been overlaid in red.

Fig. E.7: An example of annotated sequence from the time synchronized video.

direction) in the image plane.

$$P_x(\beta; C_x, M_x, \phi) = C_x - M_x * \frac{\sin(\beta)}{\cos(\phi + \beta)} \tag{E.3}$$

where ϕ is the angle between the external camera image-plane and eye-image plane, C_x is the pixel position when looking straight ($\beta = 0$), and M_x is a multiplication-factor, determining the change in pixel position with change in gaze direction. A calibration step with the user's cooperation (by asking them took in particular directions) can be performed to determine the parameters. Since the device is not firmly fixed to the head and can move during usage, we ideally need to perform calibration again. However, for our purposes we found that as long as the camera is not rotated (along the vertical-axis allowed by the device for adjusting the display), it did not degrade the performance during normal usage. A Gaussian kernel around this location is combined with the detected object based image saliency to infer the allocated attention location. This leads to the attended object as the closest object detected around the gaze-location.

4 Data set

Data is collected from naturalistic on-road driving using a vehicular test bed, which is equipped with one Google Glass and three GigE cameras as shown in fig. E.8. The

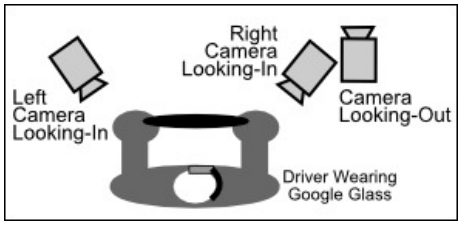


Fig. E.8: Top view diagram of the test bed setup.

Google Glass is worn by the driver to give a first-person perspective. Data is captured from Google Glass at 50 frames per second at a resolution of 1280 x 720 pixels, and stored internally on the device. Of the three GigE cameras, one is mounted to the left of the driver near the A-pillar on the windshield looking at the driver, and two are mounted to the right of driver near the rear-view mirror on the windshield - one looking at the driver and one looking outside. A multi-perspective two camera approach is adopted to look at the driver because it increases the operational range when the driver makes spatially large head movements [20]. Data from the GigE cameras is captured at a resolution of 960 x 1280 pixels and is stored on a laptop with time stamps of millisecond precision. This allows for time synchronized videos.

In order to synchronize the first-person video with the videos looking at the driver, synchronization points are annotated using the first-person view and the outside front view. The criteria used in choosing these synchronization points include naturally occurring changes in traffic lights and artificially introduced momentary but periodic bursts of light (e.g. LED lights mounted to be visible in both first person view and outside front view). Then, assuming constant frame rate in the first-person video, linear interpolation is used to synchronize the first-person video with videos looking at the driver.

Using this test bed, multiple drivers were asked to drive on local streets. Approximately 40 minutes of data was collected in total, where the drivers passed through many stop signs, traffic signals and pedestrian crossings. In this paper, we are interested in events where the vehicular test bed is near or at these intersections, because these times are especially rich with visual interaction between driver and pedestrians. To evaluate our proposed attention system, two sets of ground truth labels are created via manual annotation on interesting event segments in the driving sequences. First, we manually annotated 410 frames and 1413 pedestrians, as seen in the first person perspective camera, with bounding boxes when either their face is visible or a significant portion of their body is visible. Second, we manually annotated 300 frames of where the driver is looking in the first person view - in particular, we annotated possible pedestrian candidate(s) as shown in fig. E.7. This is accomplished by carefully looking at the driver's head and eye movements in the time synchronized videos with significant utilization of temporal and spatial context. For example, by looking at the driver's gaze over a time period, we are able to zero-in on particular pedestrians within a larger group. Annotating what the driver is looking at is especially challenging, and we have attempted to address this by obtaining consensus from multiple experts.

Table E.1: Dashboard masking cuts the false positive rate in half, without impacting the detection performance too much.

	False positives per frame (FPPF)	Detection rate
Non-filtered (baseline)	**2.94**	0.27
With dashboard filter	**1.45**	0.21

Table E.2: Performance of the attention estimator. Pedestrian accuracy shows how many pedestrian bounding boxes are correctly determined to be the attention point for the driver. Mean and median error are measures of how far the gaze-surrogate point is from the correct pedestrian bounding box.

Estimator	Mean gaze error (in pixels)	Median gaze error (in pixels)	Attended pedestrian accuracy (%)	
			Manually annotated pedestrians	Full system
Center-bias based (baseline)	148.3	127.0	55.9	37.0
Proposed	54.1	32.2	79.4	46.0

5 Experimental Evaluation

In this section we discuss the results of an experimental evaluation over several hundred frames of manually labeled data. There are two main contributions to evaluate: the impact of the dashboard masking and the attention estimation performance. In this section, both will be tested separately and then in combination.

Dashboard masking cuts the number of false positives in half with a low impact on the detection rate, as shown in table E.1. This paper is not about pedestrian detection as such, but the detection rates have been included to demonstrate that the masking does not impact them negatively in a significant way. The test set (1413 annotated pedestrians over 410 frames) is very challenging with articulated pedestrians and heavy occlusions, and while the detection numbers are low from an absolute point of view, the attention estimation still works well, as we shall see below.

The attention estimation has been tested on the same sequence with manually annotated pedestrians. Ground truth for the attentional location also was determined manually. Table E.2 shows the accuracy of the proposed system given a perfect pedestrian detector, as well as the combined system. As points of comparison, we also include results of a simple attention estimator using only head pose - the center-bias based solution. This places the focus of attention on the central field of view of the driver's head.

The gaze-surrogate estimation significantly outperforms the baseline. The extra

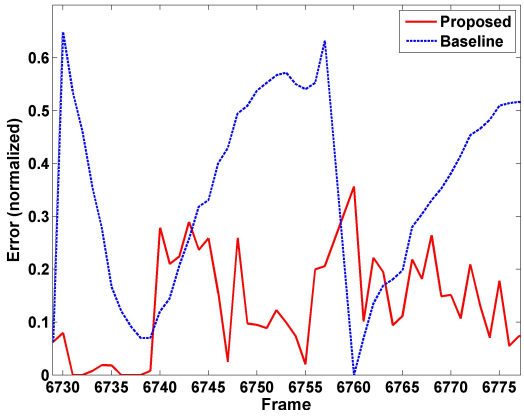


Fig. E.9: Normalized error of surrogate gaze estimate on a continuous segment.

Fig. E.10: Visualization of the LILO attention result in a sequence where gaze switches from one salient location to other. The red box around the pedestrian illustrates the salient region, the Gaussian kernel with yellow center shows the gaze location in the driver's field of view. The solid box is the pedestrian that is the subject of the driver's attention as detected by the full system.

information gained by monitoring the eye gaze on top of the head pose gives rise to much better accuracy. The full system gives the correct subject of attention in nearly half the cases and with a relatively low median error. The attention estimation works better with a perfect pedestrian detector, enhancing the accuracy by 172% from 46.0% to 79.4%. Since it relies on detected pedestrian bounding boxes, it will inevitably give the wrong output if the correct pedestrian is not detected - in that case the attention will simply be associated with the nearest detected bounding box. Implementing a perfect pedestrian detector is outside the scope of this paper, but the entire system would work with a different and better detector. It is very likely that tracking of pedestrians could improve the detection system by compensating for missed detections, but it is also worth to note that due to the, at times, rather extreme ego-motion of the driver's head, this is not a trivial task.

Fig. E.10 shows the pixel error of the gaze-surrogate detection over a full test sequence and the system is almost universally better than the baseline, except in the few situations where the subject of attention is right in the middle of the field-of-view, where the baseline system is better by sheer coincidence.

6 Concluding Remarks

We have introduced a new approach to analyzing the attention state of a human subject, given cameras focused on the subject and their environment. In particular, we are motivated by and focus on the task of analyzing the focus of attention of a human driver. We presented a Looking-In and Looking-out framework combining gaze surrogate and object based saliency to determine the focus of attention. We evaluated our system in a naturalistic real-world driving data set with no scripted experiments. This made the data set very challenging, but realistic. We showed that by combining driver state (using face analysis), we significantly improve the performance over a baseline system based on image saliency with center bias alone. The proposed framework circumvents the precise gaze estimation problem (a very challenging task in real-world environment like driving) and hence, provide a robust approach for driver focus of attention estimation.

The challenges associated with ego-centric vision are unique (with large ego-motion) and compounded by the driving environment. It presents a difficult 'in-the-wild' scenario for object detection such as pedestrians, cars etc. We propose methods to prune false detection by incorporating a region-of-interest. There is still room for improvement. In the future, we will work to provide a comprehensive and rich data set from driver's field of view camera. The novel and unique vehicle test bed will also be very useful in other areas of interest e.g. driver's activity recognition.

Bibliography

[1] A. Borji, D.N. Sihite, and L. Itti. "Probabilistic learning of task-specific visual attention". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2012, pp. 470–477. DOI: 10.1109/CVPR.2012.6247710.

[2] Ali Borji and Laurent Itti. "State-of-the-art in visual attention modeling". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.1 (2013), pp. 185–207.

[3] John Canny. "A Computational Approach to Edge Detection". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* PAMI-8.6 (Nov. 1986), pp. 679–698. ISSN: 0162-8828. DOI: 10.1109/TPAMI.1986.4767851.

[4] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*. 2005.

[5] Michael S Devyver, Akihiro Tsukada, and Takeo Kanade. "A Wearable Device for First Person Vision(FICCDAT workshop)". In: *3rd International Symposium on Quality of Life Technology*. July 2011.

[6] Anup Doshi and Mohan M. Trivedi. "Attention estimation by simultaneous observation of viewer and view". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE. 2010, pp. 21–27.

[7] Richard O. Duda and Peter E. Hart. "Use of the Hough Transformation to Detect Lines and Curves in Pictures". In: *Commun. ACM* 15.1 (Jan. 1972), pp. 11–15. ISSN: 0001-0782. DOI: 10.1145/361237.361242. URL: http://doi.acm.org/10.1145/361237.361242.

[8] Alireza Fathi, Yin Li, and JamesM. Rehg. "Learning to Recognize Daily Actions Using Gaze". In: *Computer Vision ECCV*. Ed. by Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Vol. 7572. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 314–327.

[9] L. Itti, C. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 20.11 (Nov. 1998), pp. 1254–1259. ISSN: 0162-8828. DOI: 10.1109/34.730558.

[10] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. "Learning to Predict Where Humans Look". In: *IEEE International Conference on Computer Vision (ICCV)*. 2009.

[11] T. Kanade. "First-person, inside-out vision". In: *IEEE Workshop on Egocentric Vision, CVPR*. 2009.

[12] K.M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. "Fast unsupervised ego-action learning for first-person sports videos". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2011, pp. 3241–3248. DOI: 10.1109/CVPR.2011.5995406.

[13] Zheng Lu and K. Grauman. "Story-Driven Summarization for Egocentric Video". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 2714–2721. DOI: 10.1109/CVPR.2013.350.

[14] E. Murphy-Chutorian and Mohan M. Trivedi. "Head Pose Estimation in Computer Vision: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.4 (Apr. 2009), pp. 607–626. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2008.106.

[15] K. Ogaki, K.M. Kitani, Y. Sugano, and Y. Sato. "Coupling eye-motion and ego-motion features for first-person activity recognition". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. June 2012, pp. 1–7. DOI: 10.1109/CVPRW.2012.6239188.

[16] Hamed Pirsiavash and Deva Ramanan. "Detecting Activities of Daily Living in First-person Camera Views". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012.

[17] Xiaofeng Ren and Chunhui Gu. "Figure-ground segmentation improves handled object recognition in egocentric video". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2010, pp. 3137–3144. DOI: 10.1109/CVPR.2010.5540074.

[18] B. Schiele, N. Olivier, T. Jebara, and A. Pentland. "An Interactive Computer Vision System DyPERS: Dynamic Personal Enhanced Reality System". In: *International Conference on Vision Systems*. 1999.

[19] Ekaterina H. Spriggs, Fernando De la Torre Frade, and Martial Hebert. "Temporal Segmentation and Activity Classification from First-person Sensing". In: *IEEE Workshop on Egocentric Vision, CVPR 2009*. June 2009.

[20] Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi. "Continuous Head Movement Estimator (CoHMET) for Driver Assistance: Issues, Algorithms and On-Road Evaluations". In: *IEEE Trans. Intelligent Transportation Systems* (2014).

[21] Ashish Tawari and Mohan M. Trivedi. "Dynamic Analysis of Multiple Face Videos for Robust and Continuous Estimation of Driver Gaze Zone". In: *IEEE Intelligent Vehicle Symposium* (2014).

[22] Xuehan Xiong and Fernando De la Torre Frade. "Supervised Descent Method and its Applications to Face Alignment". In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. May 2013.

Paper F

Trajectory Analysis and Prediction for Pedestrian Safety

Andreas Møgelmose, Mohan M. Trivedi, and Thomas B. Moeslund

The paper has been published in the
Proceedings of Intelligent Vehicles Symposium (IV), in press, 2015.

The layout has been revised.

Abstract

This paper presents a monocular and purely vision based pedestrian trajectory tracking and prediction framework with integrated map-based hazard inference. In Advanced Driver Assistance systems research, a lot of effort has been put into pedestrian detection over the last decade, and several pedestrian detection systems are indeed showing impressive results. Considerably less effort has been put into processing the detections further. We present a tracking system for pedestrians, which based on detection bounding boxes tracks pedestrians and is able to predict their positions in the near future.

The tracking system is combined with a module which, based on the car's GPS position acquires a map and uses the road information in the map to know where the car can drive. Then the system warns the driver about pedestrians at risk, by combining the information about hazardous areas for pedestrians with a probabilistic position prediction for all observed pedestrians.

1 Introduction

As technology advances, Advanced Driver Assistance Systems (ADAS) becomes more and more commonplace in today's cars. ADASs can range from parking assistance to safety systems such as lane departure warning, and all the way to autonomous driving in stop-and-go traffic. Our focus is on safety systems related to pedestrians. In 2009 there were 4,000 deaths and 60,000 injuries from pedestrian-vehicle collisions in the US alone[2]. Since a pedestrian is much more vulnerable than people in cars, even slow speed accidents can prove deadly.

In the scientific community, there has been a wealth of good work done on pedestrian detection[12]. The problem is not fully solved, but very reliable and fast detectors are coming out. There has been comparatively little research on what to do with these detections. In some cases, particularly autonomous cars, they can be treated as just another obstacle to avoid, but this takes no advantage of the knowledge that this particular obstacle is a pedestrian. It might as well be another car, a tree, or a trash can detected by radar. This paper is concerned with how to use pedestrian detections in a driver assistance context.

In complex driving scenarios, the driver has to take many inputs into account. Some are very important: pedestrians in front of the car, traffic signals, other road users. Some are less important: pedestrians on the sidewalk, "no parking" signs when the driver is not trying to park, billboards along the road. Even though the human mind is very good at tuning out noise, sometimes too much information is ignored, leading to accidents, e.g. from overlooking a pedestrian. This is where decision support ADASs can help. While pedestrian detectors might be able to point out every single pedestrian in a scene, that information is not very helpful for the driver, as it is too much to process.

An ADAS which aims to help in preventing pedestrian-vehicle accidents must therefore prioritize pedestrians and only inform the driver about those who are in immediate risk of being hit. Since pedestrians often move around, some predictive ability is desired for the risk assessment, as a pedestrian not in the risk zone at one

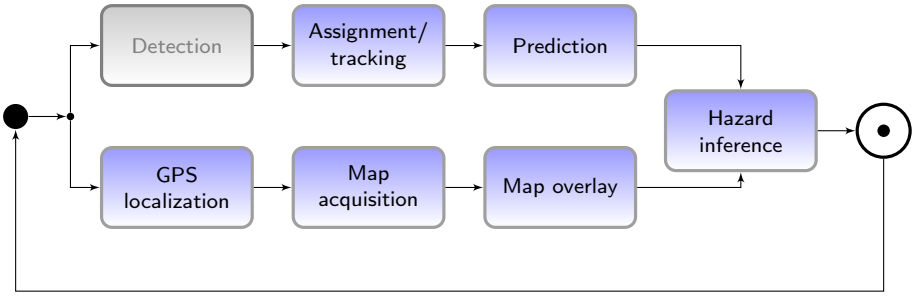


Fig. F.1: System flow. Note that *Detection* is grayed out, since it is outside of the scope of this work.

time instance might step in front of the car at the next. Being able to predict pedestrian motion will give the driver more time to react and potentially prevent an otherwise unavoidable accident.

The system presented in this paper is a monocular pedestrian tracking system. It has two parts:

1. From input of detection bounding boxes in a monocular view, it tracks pedestrians in a top-down map-like view in the area in front of the car. Using the tracks, the motion of the pedestrians in the immediate future can be predicted.
2. Based on the position and orientation of the vehicle, obtained from a highly sensitive GPS module, it retrieves a map of the area and uses the information about road locations from this to infer dangerous areas for tracked pedestrians.

The remainder of this paper is structured as follows: In section 2, a brief overview of related work is given. Section 2 gives an overview of the system, and further details are explained in section 4 and 5. Section 6 shows system output, and finally section 7 rounds off the paper.

2 Related studies

Plenty of work has been done on pedestrian detection in the past decade. The classic approaches are Haar-cascades [23] by Viola and Jones and HOG-SVM [3] by Dalal and Triggs. These two works form the foundation for much of the more recent work, from a combination of the two methods in [20], to the deformable parts model championed by Felzenszwalb et. al. [10], and Integral Channel Features[6, 5]/Aggregated Channel Features[4] by Dollár et. al. For a comprehensive overview of pedestrian detection methods, see [7].

What all these methods have in common is that they find the pedestrians, but do no further analysis. Recently, pedestrian intent prediction has gained traction, championed by the group of prof. Dariu Gavrila. There are two basic approaches: Tracking pedestrians, or looking at pedestrian orientation and local motion features, such as optical flow, of the pedestrian.

The papers [8, 11, 22] look at pedestrian orientation using different kinds of classifiers on static monocular pedestrian images. [17, 18, 9] do the same, but based on RGB-D data, and [9] even determines the orientation of the head and the torso separately. In [16], local motion featues dubbed MCHOG are used to predict whether or not the pedestrian is about to take a step. The input data, however, is not coming from a car perspective, but a stationary multi-camera setup mounted at an intersection. A similar task is carried out in the very interesting [14], which uses optical flow and stereo data. This time on data from a real car, though in rather simplistic scenarios.

Tracking of moving pedestrians is done from a surveillance perspective in [19, 1], and from a car perspective on stereo data in [21, 13] using Interacting Multiple Model Kalman Filters and SLDS tracking, respectively. Finally, long term path prediction from a stationary camera is done in [15].

The work presented in this paper also belongs to the class of tracking-based prediction systems. Its main differences to the state of the art are:

1. The system works on a monocular camera from a car perspective, where most others use RGB-D data.
2. Maps are integrated in the system and used to infer hazardous areas.
3. The analyzed scenes are complex, natural, and unconstrained with real pedestrians.
4. Particle filters are used for tracking.

3 System overview

The structure of the system presented here is shown in fig. F.1. Two parallel processes run for each frame. To begin with the upper row: Pedestrian bounding boxes are supplied from some kind of detector. Detection itself is outside the scope of this work, and throughout the project, hand-annotated bounding boxes have been used in place of a detector. Each detection in a frame is assigned to a track, or a new track is initiated if no existing track fits. Using the dynamics captured by the particle filter tracker, the pedestrian's position can be predicted into the near future.

The lower row shows the mapping part. Using a GPS and an electronic compass, the position and orientation of the ego-vehicle is determined. Based on that, a map is retrieved from OpenStreetMap and rotated to fit the surroundings of the car. A top view of the car's surroundings is generated via Inverse Perspective Mapping, and the tracks, along with the street map are superimposed to this. By using pre-acquired mapping (in this case from OpenStreetMap, but any map provider would work), we do not need to rely on road segmentation in the input images. Instead we know where the car will drive in the future.

In hazard inference, the projected position of any pedestrian is compared to the road position, and if the pedestrian enters with sufficient certainty, the driver can be warned. The actual UI for warning the driver is not covered in this paper.

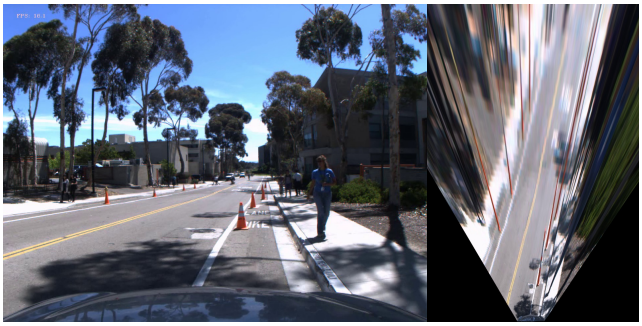


Fig. F.2: Native camera view on the left and top-down view generated via IPM on the right.

4 Trajectory generation and tracking

The trajectory generation consists of two tasks: Assignment and tracking. In assignment all detections are assigned to an existing track, or a new track is created for them. In tracking, each track is updated. The assignment takes places in the native camera view, whereas all tracking is done on the top-down map of the vehicle generated using Inverse Perspective Mapping (IPM). Fig. F.2 shows both views. The input image is a 1280x960 RGB image captured with a networked PointGrey camera.

4.1 Assignment

Assignment is done with the Munkres algorithm between bounding boxes in the current input image and the previous bounding box for each track. A cost matrix is populated with the cost for associating a bounding box with any given previous bounding box. The cost is the Euclidean distance between the box centers plus the size change of the box (a bounding box is expected to be roughly the same size in two consecutive frames). Since boxes move and change size in bigger increments when pedestrians are close to the camera, the cost is weighted by the inverse of the box size, so large costs are lowered when the bounding box is large:

$$D(a,b) = \left(\sqrt{|a_x - b_x|^2 + |a_y - b_y|^2} + \sqrt{|a_w - b_w|^2 + |a_h - b_h|^2} \right) \cdot \frac{1}{a_w + a_h} \quad \text{(F.1)}$$

where $D(a,b)$ is the distance between boxes a and b, and subscripts x, y, w, h means center x-coordinate, center y-coordinate, width, and height, respectively.

10 bounding boxes gives a 10×10 matrix. The Munkres algorithm allows for unequal numbers of input and output by padding the cost matrix with "infinity" until it is square. That way, assignments can be still be made, and the leftover in-/outputs are simply assigned to nothing. This, however, finds a global optimum, and will often lead to all boxes jumping around. Imagine a case where one track ends, and another begins simultaneously. There will still be an equal number of boxes on the in- and output side, so all boxes will be reassigned, when in reality one box should have been assigned to nothing and one should have prompted a new track. To accommodate

Fig. F.3: Example of assignment between pedestrians. Two boxes are shown per person: the current bounding box and the previous bounding box from the track the pedestrian is assigned to. Some boxes are very close to each other, and thus hard to distinguish on the picture.

these scenarios, we add another 10 columns of close-costs. The close-cost is simply a threshold over which we decide that it is better to close a track than to reassign it. As a result, in the 10×10 case, the cost matrix will now be 10×20, with the left half containing proper reassignment costs, and the entire right half containing indentical close-costs in all entries.

After the assignment in done, any unassigned detection will be assigned a new track. An assignment example is shown in fig. F.3.

4.2 Tracking

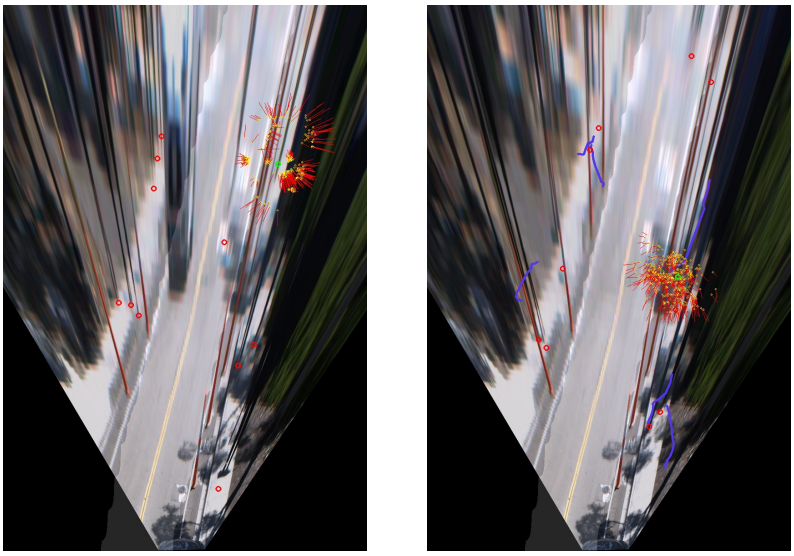
In this system tracking is done by particle filter. In the following a single filter is described, but each track has its own. Each track has 1000 particles that are modeled using the unicycle model: a particle has a certain speed and orientation. Each update is done by applying a certain amount of Gaussian noise to each of these parameters. Each measurement is also applied Gaussian measurement noise. The particles are weighted by distance to the measurement using a bivariate Gaussian:

$$P(p,m) = \exp\left(\frac{(p_x - m_x)^2}{2 \cdot v_x} - \frac{(p_y - m_y)^2}{2 \cdot v_y}\right) \tag{F.2}$$

where $P(p,m)$ is the probability of a particle p given the measurement m and v is the variance in two dimensions. Example pictures of the particle filter tracking are shown in fig. F.4

5 Behavior prediction and hazard inference

The prediction is carried out by the particle filter. When the prediction has been computed for the desired prediction horizon, a bivariate Gaussian is fitted over the particles. While particle filters support multimodal hypotheses, that is practically never seen with a relatively direct measurement setup as in this system, so fitting a single bivariate Gaussian does not lead to a significant loss of information. The fitted Gaussian is then used to describe the probability of positions the pedestrian might be at in the near future. Fig. F.5a visualizes the probability map one step ahead and fig. F.5b shows the prediction 5 estimates ahead.



(a) (b)

Fig. F.4: All 1000 particles for a single track (the green observation) plotted with position and orientation. (a) shows the distribution before convergence and (b) shows the distribution after.

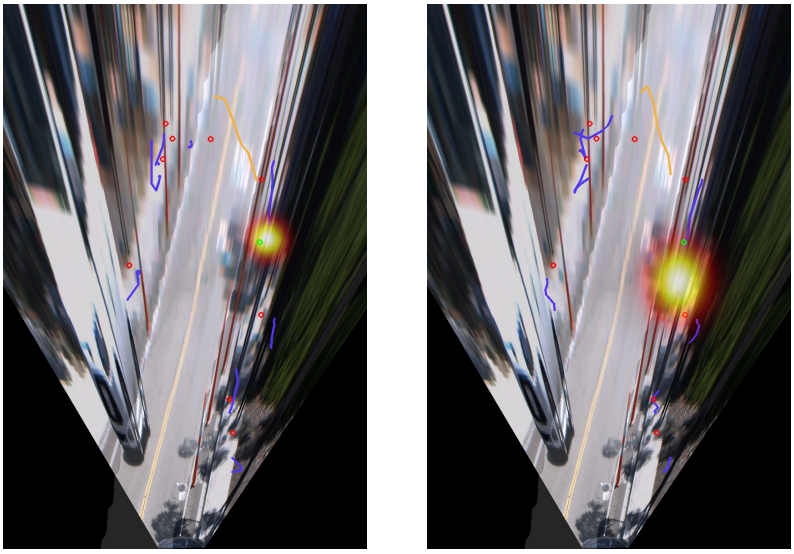
A central step in the hazard inference process is using a map to determine where the road is. Pedestrians on the sidewalk are close to the road, but a perfectly normal sight and not automatically a hazard. To determine when pedestrians are about to enter a dangerous zone, it is imperative to know where the road is. By using a map, the system does not have to rely on road segmentation or vehicle dynamics. Road segmentation is a hard problem to solve reliably, and while vehicle dynamics are useful in short-term collision avoidance, it is of less use in slow driving scenarios or complex situations with sudden changes in orientation, such as roundabouts.

A map is obtained from OpenStreetMap, which does not necessarily require internet access, but can be installed on a server locally in the car. Any other mapping source can also be used. Then the map is rotated according to the car's orientation and scaled appropriately. Fig. F.6 illustrates this process.

Now, to estimate whether a pedestrian is about to enter a hazardous zone, a combination of the predicted position – as visualized by the heatmaps in fig. F.5 – and the known road area is used. Each pixel of the heatmap has a value, depending on the probability of the pedestrian being there. The values of the pixels overlapping with the road are summed, and if the sum is above some threshold, a warning is emitted:

$$W(p) = \sum_{(x,y)\in H} H(x,y) \cdot R(x,y) \tag{F.3}$$

where $W(p)$ is the warning-value (subject to a threshold) of particle p, H is the position probability map of p and R is the mask of the road, expressed as 1 where road is present and 0 otherwise.



(a) (b)

Fig. F.5: Example position predictions of the green observation (a) one time step ahead and (b) five steps ahead. The one-step prediction is behind the actual observation because of measurement-noise, which makes the observation unreliable and the filter smooths the movement out. This is exacerbated by the inverse perspective mapping, which is very sensitive to even small changes in bounding box position, especially at a distance. As the prediction is done further into the future, the uncertainty rises, which can be seen by the expanding heatmap.

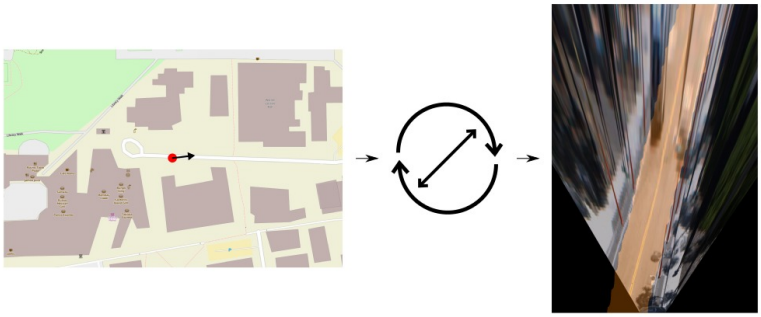


Fig. F.6: The road in front of the car is extracted from OpenStreetMap. It is then rotated and scaled appropriately and overlaid on the IPM-generated map.

6 Evaluation and discussion

Figure F.7 shows output examples of the full system. The left side show the native view with detection bounding boxes, while the right side shows the map view with circles for pedestrians. Heatmaps on the map view show the predicted position of the pedestrians, and the color of both bounding boxes and circles show whether the pedestrian is about to enter a hazardous area.

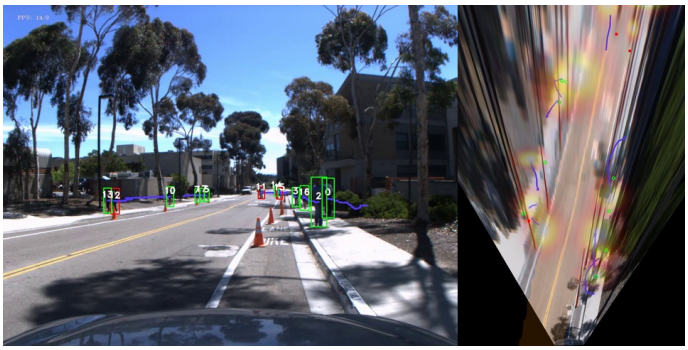
Assignment: The system consists of several blocks, and in this section we discuss the performance of each. The first is assignment, where each bounding box is associated with a track. This part performs very well. On a test sequence with 1022 individual detections over 106 frames, only 4 wrong assignments happen. Two of these instances are shown in fig. F.8. In the top row, the person of track 13 is exiting the frame. In F.8b, after 13 having exited, the person previously assigned to track 12 steps to the exact position that 13 had before, and the algorithm determines that the lowest global cost is achieved by closing track 12. This issue might be solved by having a stronger prediction input to the assignment.

In F.8c and F.8d, the previous track 3 is occluded by a person in front of him at the exact same time as a pedestrian which was previously hidden emerges in the same line-of-sight of the camera. The two boxes are practically in the same position - though with some size change - and thus the assignment cost is sufficiently low for the wrong assignment to happen.

Tracking: The tracking has two jobs. Smooth out the trajectory of the tracked pedestrians and allow for prediction of their position. The smoothing is especially

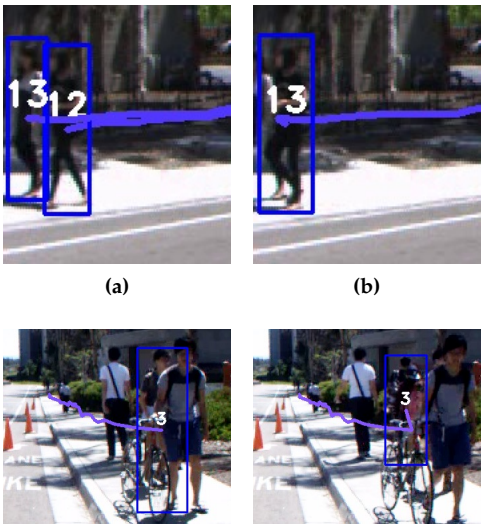
(a)

(b)



(c)

Fig. F.7: Hazard warnings at different times. Green boxes/circles are pedestrians who are not about to enter the road. Red indicates pedestrians who either are on the road already, or are about to enter it. (a) is very early in the sequence, so the tracks are still uncertain, and the heatmaps are large. The remaining examples are taken later, when the tracks are more reliable. (c) has a faulty warning at track 12, due to a very short track with great uncertainty (the previous tracks were cut when a car passed). It also has a missing warning all the way in the background at track 14. This is due to that pedestrian being so far away that he is outside of the map, and thus not considered for the risk analysis.



(a) (b)

(c) (d)

Fig. F.8: Two of the just four wrong assignments in the 1022 detections of the test sequence. Each row is an example with the before and after frame shown. See text for further discussion.

necessary in a monocular setup, such as ours, where the distance to pedestrians is detemined via IPM. At far distances, the resolution of the camera limits the depth resolution to be very coarse. Thus, if a bounding box position differs by even just a pixel in the y-axis, the estimated position on the map can jump several meters. The tracking should compensate for this effect and give a smoother trajectory closer to the way pedestrians actually move. Setting up exact metrics for this is difficult at best, when the ground truth position of the pedestrians is not known.

An example of this is shown in fig. F.9. Here, the original input is shown in blue, and the particle filter output in orange. The smoothing can be adjusted via the system noise until a satisfying combination of smoothness and reaction time is reached. From a visual inspection it is clear that the orange tracks provide a much better approximation to the real world than the very jagged input data, even at relatively long distances. This is especially clear in the tracks to the left.

Hazard inference: The final block is the hazard inference which combines the predictive power of the tracking with the map based road localization. Hazard inference examples are shown in fig. F.10. F.10a shows a successful - but easy - prediction, where a crossing cyclist is in the middle of the road and is not predicted to leave the road in the next 2 seconds. F.10b shows the situation a few frames later, as the cyclist is just about to leave the road, and is thus predicted to be out of danger soon. F.10c shows another use for the system. Here, there is no prediction of impending hazards, but there is clearly a large concentration of pedestrian activity in the lower right corner. This knowledge in itself might be useful in a driver assistance context. Finally, F.10d shows a faulty prediction. In this case a pedestrian steps onto the bike

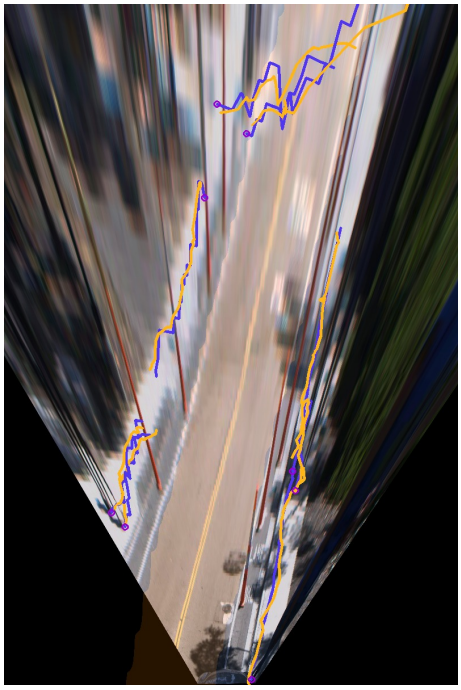


Fig. F.9: Example of how the tracking (orange lines) is smoothing the original input (blue lines).

lane to overtake slower pedestrians on the sidewalk. The system predict her to enter the roadway, but in fact she never does. Technically, this makes it a false prediction, but it could be argued that it is beneficial to warn the driver about this still, since she is in very close proximity to the road.

7 Concluding remarks

This work presented a pedestrian intent prediction system for use in driver assistance. It uses monocular a monocular view of the road, in which detected pedestrians are mapped to a top view computed using Inverse Perspective Mapping. For tracking, pedestrians are assigned to a track based on their bounding box in the native camera view, and tracking takes place in the map view using particle filters.

To determine which areas in view are potential hazardous zones for pedestrian, external mapping is used. A map of the nearby area is acquired from OpenStreetMap and superimposed onto the IPM view. Using the knowledge of road positions in the map, hazardous areas in the camera view are obtained. Based on the overlap of the trackers' predictions for all observed pedestrians with the road area, the driver can be informed about wayward pedestrians.

In the future, ego-motion compensation should be added to the system, so it works for moving vehicles. Furthermore, the orientation of the pedestrians – based

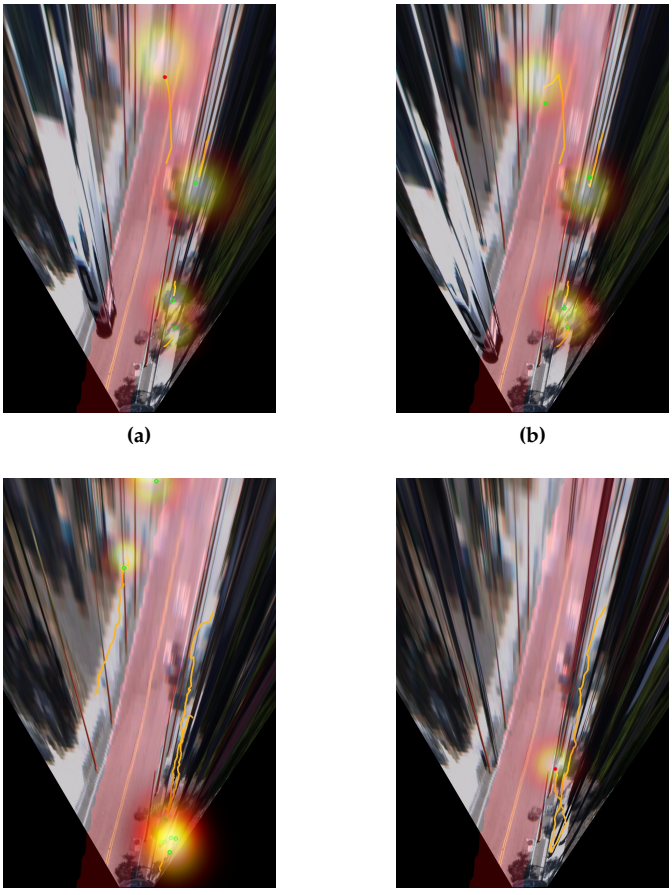


Fig. F.10: Hazard inference in 4 situations. The hazardous area - the road - is marked with a translucent red color. For clarity, only a few tracks are shown in each. (a) and (b) are just a few frames apart, and the hazard indicator for the subject crossing the road goes from red to green when he is predicted to leave the road. (c) shows an example of heavy pedestrian activity, and (d) shows a mistaken prediction.

on appearance, not dynamics – should be included in the tracking measurements, since pedestrians are capable of very rapid orientation changes, which are hard to capture in a purely dynamics-based system such as this. It is also possible that local motion cues from e.g. the pedestrians' legs can be used in improving performance.

Acknowledgment

The authors would like to thank their colleagues at the LISA lab for valuable discussions throughout the project.

Bibliography

[1] Aniket Bera, Nico Galoppo, Dillon Sharlet, Adam Lake, and Dinesh Manocha. "Adapt: real-time adaptive pedestrian tracking for crowded scenes". In: *Proceedings of Conference on Robotics and Automation, Hong Kong*. 2014.

[2] Jonathan Cinnamon, Nadine Schuurman, and S Morad Hameed. "Pedestrian injury and human behaviour: observing road-rule violations at high-incident intersections". In: *PloS one* 6.6 (2011), e21063.

[3] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*. 2005.

[4] P. Dollar, R. Appel, S. Belongie, and P. Perona. "Fast Feature Pyramids for Object Detection". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.8 (Aug. 2014), pp. 1532–1545. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2300479.

[5] Piotr Dollár, Serge Belongie, and Pietro Perona. "The Fastest Pedestrian Detector in the West". In: *BMVC*. Vol. 2. 3. Citeseer. 2010, p. 7.

[6] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. "Integral Channel Features". In: *BMVC*. Vol. 2. 3. 2009, p. 5.

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. "Pedestrian Detection: An Evaluation of the State of the Art". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.4 (2012), pp. 743–761.

[8] M. Enzweiler and D.M. Gavrila. "Integrated pedestrian classification and orientation estimation". In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. June 2010, pp. 982–989. DOI: 10.1109/CVPR.2010.5540110.

[9] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila. "Joint probabilistic pedestrian head and body orientation estimation". In: *Intelligent Vehicles Symposium (IV), 2014 IEEE*. June 2014, pp. 617–622.

[10] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. "Object Detection with Discriminatively Trained Part-Based Models." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32.9 (2010), pp. 1627–1645.

[11] T. Gandhi and Mohan M. Trivedi. "Image based estimation of pedestrian orientation for improving path prediction". In: *Intelligent Vehicles Symposium, 2008 IEEE*. June 2008, pp. 506–511. DOI: 10.1109/IVS.2008.4621257.

[12] Tarak Gandhi and Mohan M. Trivedi. "Pedestrian protection systems: Issues, survey, and challenges". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.3 (2007), pp. 413–430.

[13] J. F. P. Kooij, N. Schneider, and D. M. Gavrila. "Analysis of pedestrian dynamics from a vehicle perspective". In: *Intelligent Vehicles Symposium (IV), 2014 IEEE*. June 2014, pp. 1445–1450.

[14] C.G. Keller and D.M. Gavrila. "Will the Pedestrian Cross? A Study on Pedestrian Path Prediction". In: *Intelligent Transportation Systems, IEEE Transactions on* 15.2 (Apr. 2014), pp. 494–506.

[15] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. "Activity forecasting". In: *Computer Vision–ECCV 2012*. Springer, 2012, pp. 201–214.

[16] S. Köhler, M. Goldhammer, S. Bauer, S. Zecha, K. Doll, U. Brunsmann, and K. Dietmayer. "Stationary Detection of the Pedestrian's Intention at Intersections". In: *Intelligent Transportation Systems Magazine, IEEE* 5.4 (winter 2013), pp. 87–99. ISSN: 1939-1390. DOI: 10.1109/MITS.2013.2276939.

[17] M. C. Liem and D. M. Gavrila. "Person appearance modeling and orientation estimation using Spherical Harmonics". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2013), pp. 1–6.

[18] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li. "Accurate Estimation of Human Body Orientation From RGB-D Sensors". In: *Cybernetics, IEEE Transactions on* 43.5 (Oct. 2013), pp. 1442–1452. ISSN: 2168-2267.

[19] Brendan Tran Morris and Mohan M. Trivedi. "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach". In: *Transactions on Pattern Analysis and Machine Intelligence* 33.11 (2011), pp. 2287–2301.

[20] Antonio Prioletti, Andreas Møgelmose, Paolo Grisleri, Mohan M. Trivedi, Alberto Broggi, and Thomas B. Moeslund. "Part-Based Pedestrian Detection and Feature-Based Tracking for Driver Assistance: Real-Time, Robust Algorithms, and Evaluation." In: *IEEE Transactions on Intelligent Transportation Systems* 14.3 (Sept. 2013), pp. 1346–1359.

[21] Nicolas Schneider and Dariu M. Gavrila. "Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study". In: *Pattern Recognition*. Ed. by Joachim Weickert, Matthias Hein, and Bernt Schiele. Vol. 8142. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 174–183. ISBN: 978-3-642-40601-0.

[22] Junli Tao and Reinhard Klette. "Integrated Pedestrian and Direction Classification Using a Random Decision Forest". In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. Dec. 2013.

[23] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* 1 (2001), pp. 511–518. ISSN: 1063-6919.

Part IV

Person Re-identification

Paper G

Multimodal Person Re-identification Using RGB-D Sensors and a Transient Identification Database

Andreas Møgelmose, Thomas B. Moeslund, and Kamal Nasrollahi

The paper has been published in the
Proceedings of the 1st International Workshop on Biometrics and Forensics (IWBF),
pp. 1–4, 2013.

The layout has been revised.

Abstract

This paper describes a system for person re-identification using RGB-D sensors. The system covers the full flow, from detection of subjects, over contour extraction, to re-identification using soft biometrics. The biometrics in question are part-based color histograms and the subjects height. Subjects are added to a transient database and re-identified based on the distance between recorded biometrics and the currently measured metrics. The system works on live video and requires no collaboration from the subjects. The system achieves a 68% re-identification rate with no wrong re-identifications, a result that compares favorable with commercial systems as well as other very recent multimodal re-identification systems.

1 Introduction

Person re-identification is useful in many contexts, and can be used as a forensics tool in most situations where surveillance cameras has captured an incident. Re-identification is the act of recognizing persons entering a camera's field of view and have been seen previously by a different camera, or by the same camera at a different time instance. The crucial difference between this and tracking is that for re-identification there is expected to be a significant spatial or temporal difference between observations, making it impossible to rely on simple motion dynamics as tracking often does. Instead, soft biometrics are used to decide if a subject has been seen before.

A number of challenges and characteristics set re-identification apart from traditional tracking and hard-biometric recognition:

- The set of re-identifiable persons must be updated on the fly; there can be no enrollment phase that requires direct participation from the subjects.
- There is no – or only weak – constraints on the pose of subjects, so the system must be robust to pose changes.
- Persons must be re-identifiable at distances where sensor resolution is generally not sufficient for traditional face recognition.
- The database containing the subjects has a transient nature since subjects are generally not relevant if they have not been re-identified after a certain time span – then they have probably left the area.

Some applications of re-identification does not require all recorded persons to be re-identified. An example is the commercial system from Blip Systems [3], which does person flow tracking in airports based on radio signatures from mobile phones. It has a re-identification rate of around 10%, which is sufficient for a representative flow map.

Because the re-identification scenario can be harder than traditional recognition due to the worse data quality, it is an obvious idea to use more sensor modalities. With the advent of the Microsoft Kinect and similar structured light-based sensors (ASUS Xtion and the PrimeSense Sensor), RGB-D sensors have become much more

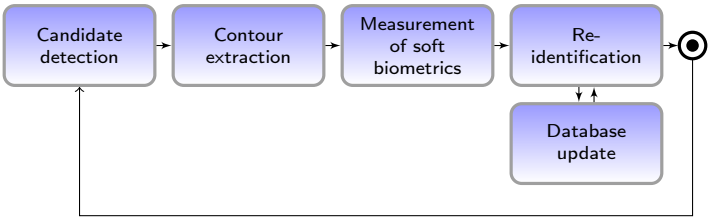


Fig. G.1: The flow of the method discussed in this paper. The loop runs once per frame.

accessible and affordable, and using them in larger surveillance applications does not seem impossible. While sensors relying on structured light have some issues, especially with outdoor use, we believe that in the near future, many more modalities – such as depth – will be integrated in surveillance cameras, and as such it is prudent for the surveillance computer vision research community to direct its attention toward multimodal methods.

The main contribution of this paper is a RGB-D based re-identification system. It performs all the steps necessary in these kinds of systems: Person detection, measurement of soft biometrics, forming and maintaining a transient database of subjects, and re-identifying subjects, whereas previous RGB-D based contributions (see section 2) has only covered parts of the process.

The rest of the paper is structured as follows: Section 2 takes a brief look at related work in the area of re-identification. Section 3 describes the structure of the re-identification methods used in the system, and contains subsections going in further details with each step. The transient database is treated in section 4, which is followed by experiments and tests of the system in section 5. The closing remarks can be found in section 6.

2 Related work

Re-identification is a relatively young field, and most contributions so far are based on regular camera input. Notable examples include [1, 12, 6, 11, 9].

Re-identification using RGB-D sensors is still in its infancy; only a few papers on RGB-D re-identification exist. [2] present a re-identification method based solely on depth-features using several normalized measures of body parts, calculated from joint positions. They include measures of the body's "roundness", which can act as a crude proxy for volume. This, however requires a high depth resolution, and is only suitable when subjects are close to the sensor. The paper is focused solely on the re-identification step and does not treat identification or extraction of joints, while our paper presents a full system. Another approach for re-identification using soft biometrics was put forth in [10], but they use manual measurements instead of automatic analysis of RGB-D images.

3 Method overview

This system covers all stages through a complete re-identification flow. An overview is presented in figure G.1. First thing to happen is the candidate detection: Before any re-identification can take place, it is necessary to know if - and where - the person is. Next step is contour extraction. A detector usually only returns a bounding box, and possibly even a bounding box that does not fit closely around the person. When extracting the biometrics, it is important not to extract information from the background, but from the subject alone, since the subject will appear on different backgrounds later and must be re-identifiable then. Next step is measurement of the desired soft biometrics to generate a descriptor for this particular subject. Finally, the candidate is either identified, added to a database of previously seen persons, or ignored due to too low data quality (this particular situation is covered in further depth in section 4).

3.1 Candidate detection

The detection stage consists of a state-of-the-art HOG-SVM detector trained on the INRIA dataset [5] run on the RGB image.

Since re-identification is not tracking, continuous detection in every frame is not necessary. As long as one good detection - or however many the re-identification process takes - is present, the re-identification can be performed. In the context of flow analysis, it is also enough to detect and re-identify the 10% of subjects with the strongest response (as done by [3]). Because of this, the detection rate can be lowered, to the benefit of the false detection rate.

3.2 Contour extraction

The depth image is used for contour extraction. When seen from a distance, persons in the depth image are generally a plane with only small depth variations, so a flood fill is used. A seed point is selected in the middle of the chest of the detected person. The starting point is defined as

$$(x, y) = \left(x_b + \frac{w_b}{2}, y_b + \frac{h_b}{3} \right) \tag{G.1}$$

where x_b and y_b are the coordinates of the upper left corner of the detection bounding box, and w_b and h_b are the width and height, respectively.

A problem arises when the fill reaches the floor. A line on the floor will be in the same depth as the feet, and thus the fill continues onto the floor. To counter this, during the initialization of the system, the ground plane is estimated. This is done by manually clicking a number of points that are on the floor and fitting a plane through these using a least squares fit performed by SVD factorization[8]. Pixels in the depth image that are near the estimated ground plane are discarded from the fill.

Fig. G.2: A real subject split in parts. The histograms are calculated based on pixel values inside the shaded areas in the boxes. RGB image on the left, depth image on the right.

3.3 Measurement of soft biometrics

Two soft biometrics are used in the system: A part-based color histogram and the subject's height. The height is found by subtracting the y-values in world-coordinates for the up most and the lowest point in a contour.

The histogram is calculated on parts of the subject to account for the differing colors between leg garments and jacket/shirt. According to [7], the legs occupy 0% to 55% of the full body height and the torso from 55% to 84% of the full height. This implementation takes its base in these figures, but since the division between legs and torso can vary from person to person, an undefined zone is introduced at the middle of the body, which is not counted in either histogram. Thus, the division used here can be seen in fig. G.2.

For each part, a histogram is created for each of the R, G, and B channels with 20 bins. These are concatenated for at total of 60 bins per part. Then the part histograms are also concatenated and the full 120 bin histogram is normalized so all bins sum to 1. This makes sure that comparison across different sizes is possible, and evens out changes in lighting.

Because this system runs on real-world data and the segmentation becomes unstable at far distances, only subjects within 4 m of the camera are considered.

3.4 Re-identification

The re-identification step consists of a comparison of the candidate to the persons saved in the transient database. The details on the database can be found in the next section, but the most basic functionality is comparison of the histograms. This is done using the Bhattachariyya distance [4]:

$$d(H_1, H_2) = \sqrt{1 - \sum_I \frac{\sqrt{H_1(I)H_2(I)}}{\sqrt{\sum_I H_1(I) \cdot \sum_I H_2(I)}}} \tag{G.2}$$

where $d(H_1, H_2)$ is the distance between the histograms H_1 and H_2, $\bar{H}$ denotes the norm of a histogram, N is the number of bins, and $H(I)$ is the value of bin I in the histogram H. The distance is a number between 0 and 1, where 0 is a perfect match.

The next section describes how the height and the distance between histograms are used with the database.

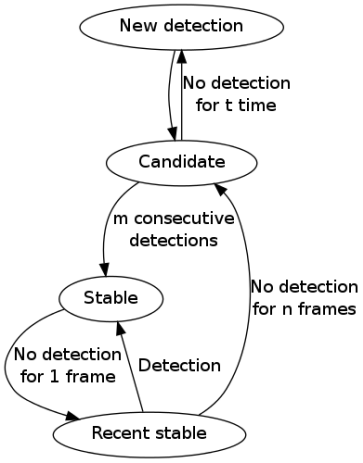


Fig. G.3: State diagram for the transient database.

4 Transient database

The purpose of the transient database is to contain the previously detected persons. Because subjects are often only seen briefly, there are very few samples per person and they are often not very structured. This makes the creation of a parametric model unreliable. Instead, we model each person with all the previous heights and histograms that have been connected to her. This gives a broad model of each candidate in many poses, orientations, and sizes. A person is then re-identified if the query biometrics are sufficiently close to either of the existing samples in a database entry for one person. In essence this is a box classifier with the histogram distance and the height as features. A box classifier makes sense since it is a reasonable assumption that the height and the clothing color are not correlated.

To minimize noise, some temporal constraints must be fulfilled before a person is considered as recognized. Thus, the detected subjects can have several states (see fig. G.3):

- New detection
- Candidate
- Stable
- Recently stable

New detection is the initial state for any new detection. If the subject is sufficiently dissimilar to the existing database entries, she is added to the database as a *candidate*. Because there is always a risk of false detections, the candidate must have been detected for at least m consecutive frames in order to become *stable*. Only stable persons are considered for re-identification.

If a stable person is not successfully re-identified, she is transferred to *recently stable*, which allows her to regain stable-state from just a single detection. If a person

has been recently stable for n frames, she is transferred to the candidate stage and must regain stability on equal terms with all other candidates. Finally, a candidate that has not been seen for a given time measured in frames, p, is likely to have left the venue, and is thus discarded completely from the database. This ensures that the database size remains at a size where it is feasible to search for new candidates quickly.

5 Experiments

A set of recordings was used to evaluate the system. They contain 25 subjects which walk past the camera twice each (not in any particular order). In total the test set consists of 7800 frames. The results are presented in table G.1. They were counted on each pass of a subject, so one pass with a correct re-identification counts for 1 in that category.

Correct re-identifications are exactly that: The subject is identified with a correct label. Ambiguous re-identifications are instances where the subject is re-identified as several people during a pass, but at least one of them is the correct label. They are still only detected as a single person, but the identification differs from frame to frame. Not enrolled means that a person is not re-identified due to the fact that the first pass did not result in any sufficiently good features, so she was never enrolled in the database. Not re-identified means that a person was enrolled, but not recognized in the subsequent pass. Finally, wrong re-identification describes the case where a person is erroneously classified with the label of another person.

76% of the subjects are correctly identified (albeit with 8% ambiguously re-identified), 20% was not identified due to missing enrollment on their first pass, and a single person was not re-identified, even though she was correctly enrolled. A design choice was to ensure that wrong re-identifications would not occur. This has been achieved successfully, but it of course has an adverse effect on the re-identification percentage. It should be noted, however, that a successful re-identification of 76% is significantly better than what [3] are capable of for their commercial system.

A very recent RGB-D re-identification study by Barbosa et. al. [2] achieves a rank 1 re-identification rate of around 15%, significantly worse than our results. In this context, rank relates to the confidence with which the person is re-identified. When a subject is re-identified, the system makes a ranked list of likely labels. Rank 1 means that the correct label is the highest ranked. Rank 5 would mean that the correct label is within the 5 highest ranked labels. Thus, their rank 1 result of around 15% is the number that can be directly compared with our results of 76%. They report their main results as the normalized area under the Cumulated Match Characteristics (CMC) curve where the x-axis is the rank and the y-axis is re-identification rate. Since our system is intended for fully automated use, anything but the rank 1 result is largely irrelevant, but if the system were to be used in a supervised context, the nAUC is indeed interesting.

Table G.1: Experimental results

Subjects	Absolute	%
Correct re-identifications	17	68%
Ambiguous re-identifications	2	8%
Not enrolled	5	20%
Not re-identified	1	4%
Wrong re-identifications	0	0%

6 Concluding remarks

This paper presented a full RGB-D based person re-identification system. On our test set, it achieves a re-identification rate of 68%. This outperforms both commercial systems for person flow tracking [3] and very recent multimodal systems [2]. It works on real-world data and covers the full system from detection through contour extraction, measurement of soft biometrics and the actual re-identification. The system exhibits a weakness with similarly dressed persons, something that might be solved be adding more advanced biometrics. Tracking might improve performance. Moreover, like others, the system assumes no occlusions, something that should be addressed in future work.

Bibliography

[1] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. "Learning to Match Appearances by Correlations in a Covariance Metric Space". In: *ECCV (3)*. Vol. 7574. LNCS. Springer, 2012, pp. 806–820. ISBN: 978-3-642-33711-6.

[2] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. "Re-identification with RGB-D Sensors." In: *ECCV Workshops (1)*. Vol. 7583. LNCS. Springer, 2012, pp. 433–442. ISBN: 978-3-642-33862-5.

[3] Blip Systems. *Blip Track Airport*. http://www.bliptrack.com/airport/area-of-operations/. 2012.

[4] G. Bradski and A. Kaehler. "Learning OpenCV". In: O'Reilly, 2008. Chap. 7, pp. 201–202.

[5] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*. 2005.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. "Person re-identification by symmetry-driven accumulation of local features". In: *CVPR*. 2010.

[7] P. Fihl, R. Corlin, S. Park, Thomas B. Moeslund, and Mohan M. Trivedi. "Tracking of Individuals in Very Long Video Sequences." In: *ISVC (1)*. Vol. 4291. LNCS. Springer, 2006, pp. 60–69. ISBN: 3-540-48628-3.

[8] G.H. Golub and C. Reinsch. "Singular value decomposition and least squares solutions". English. In: *Numerische Mathematik* 14 (5 1970), pp. 403–420. ISSN: 0029-599X. DOI: 10.1007/BF02163027.

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. "Person Re-identification by Descriptive and Discriminative Classification". In: *SCIA*. Vol. 6688. Lecture Notes in Computer Science. Springer, 2011, pp. 91–102. ISBN: 978-3-642-21226-0.

[10] C. Velardo and J. Dugelay. "Improving Identification by Pruning: A Case Study on Face Recognition and Body Soft Biometric". In: *WIAMIS*. IEEE, 2012, pp. 1–4. ISBN: 978-1-4673-0791-8.

[11] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. “Shape and Appearance Context Modeling”. In: *ICCV*. IEEE, 2007, pp. 1–8.

[12] W. Zheng, S. Gong, and T. Xiang. “Person re-identification by probabilistic relative distance comparison”. In: *CVPR*. IEEE, 2011, pp. 649–656.

Paper H

Comparison of Multi-shot Models for Short-term Re-identification of People using RGB-D Sensors

Andreas Møgelmose, Chris Bahnsen, and Thomas B. Moeslund

The paper has been published in the
Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), 2015.

The layout has been revised.

Abstract

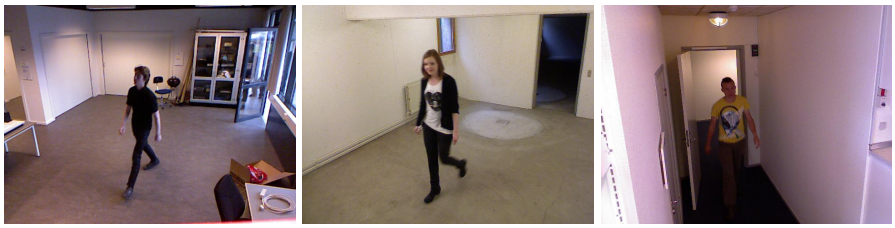
This work explores different types of multi-shot descriptors for re-identification in an on-the-fly enrolled environment using RGB-D sensors. We present a full re-identification pipeline complete with detection, segmentation, feature extraction, and re-identification, which expands on previous work by using multi-shot descriptors modeling people over a full camera pass instead of single frames with no temporal linking. We compare two different multi-shot models; mean histogram and histogram series, and test them each in 3 different color spaces. Both histogram descriptors are assisted by a depth-based pruning step where unlikely candidates are filtered away. Tests are run on 3 sequences captured in different circumstances and lighting situations to ensure proper generalization and lighting/environment invariance.

1 Introduction

The task of person re-identification is about recognizing people that have been captured earlier by a camera in a surveillance network. The network may consist of one or more cameras, and can be placed in traditional surveillance contexts or more narrowly scoped areas, such as keeping track of a single queue of people. The objective is simple: When a person enters the field of view of a camera in the system, it must be determined whether or not this person has been seen before. Person re-identification is closely related to person tracking and person recognition. However, is has several extra challenges, that makes it less straight-forward [9]:

- There is no fully known gallery dataset. As opposed to traditional person recognition, the system must enroll new people on-the-fly, without them taking any action.
- Methods must be robust to pose changes. Since subjects are not required to participate actively, there are only weak constraints on pose and viewing angles.
- Sensor resolution is a big challenge. People simply passing by at various distances are to be re-identified, so it is not reasonable to use hard biometrics like fingerprints or face recognition.
- The database of known people must be continually cleaned up - when a person has not been seen for some period of time, they have most likely left the area and should be removed from the database.

There are two fundamentally different approaches to re-identification: Single-shot and multi-shot. Single-shot performs the re-identification on stand-alone frames. This is useful in situations where only a single probe image is available. However, very often the subject has been captured on video, and thus has several frames describing her. Multi-shot combines a full pass across the field of view into a single model, which is the used as probe in a gallery of similarly collected multi-shot models. Multi-shot gives the option of capturing more information about the subject than a single frame contains, and has the potential to make the system more robust to occlusions and sudden changes in lighting.



(a) (b) (c)

Fig. H.1: Example images from our own (a) Novi, (b) Basement, and (c) Hallway sequences.

Person re-identification has been in active research for a while, but multi-modal systems have only recently come into play. The reason for this is twofold: 1) Algorithms have so far mostly been developed for use in existing surveillance infrastructure and 2) more advanced sensor capabilities, such as depth and thermal, have not been readily available. We believe that as sensor technology progresses, more modalities will show up in regular surveillance cameras, making the development of new multi-modal algorithms highly relevant.

This work builds on the method presented in [9] and is a full RGB-D based re-identification system covering all parts of the pipeline from detection through re-identification to database maintenance. The main contributions are:

- While the earlier work was single-shot based, the method has been updated to a multi-shot approach. This work compares several different multi-shot person models.
- The earlier work relied on RGB-color histograms. This work presents a comparison of three different color spaces: RGB, HSV, and XYZ.
- More thorough testing. On top of testing on the original dataset from [9], two more datasets have been captured to test the performance in different circumstances.
- The system is now free of arbitrary thresholds in the re-identification stage, as every threshold is learned from training data in a cross-validation scheme.
- In the original work, the height of subjects only had little influence on the re-id performance. We introduce a more thorough pruning step based on depth-adjusted height of subjects which increases re-id performance significantly.

The remainder of this paper is structured as follows: Section 2 gives an overview of related work in the field of re-identification. It also contains a description of existing datasets, as well as the ones captured and used in this work. Section 3 explains the algorithms used and goes through detection and segmentation, multi-shot person modeling, and re-identification. In section 4 the various methods presented are evaluated against each other. Section 5 concludes the paper.

2 Related work

Person re-identification as described above has been an active research area for about a decade and truly gained speed in the latter half of the 2000s. A relatively recent survey on person re-identification can be found in [6], and in this section we highlight notable recent papers. As mentioned previously, re-identification approaches can be divided into single-shot and multi-shot. Furthermore, we distinguish whether multi-modal methods are used.

Zheng et. al. [12] and Zhao et. al. [11] both use single shot algorithms. The first use color and texture histograms, whereas the latter uses dense color histograms and SIFT descriptors with the addition of using a saliency map to decide which parts of the person are the most descriptive.

Multi-shot is championed by Bak et. al. in [1] and Demirkus et. al. [5]. Bak uses a large pool of features and the best one to describe a particular person is selected. Demirkus uses a set of more directly understandable soft biometrics, such as gender, hair color, and clothing color.

Moving away from the traditional visible light modality, Jüngling and Arens [7], presents a full single-shot re-identification pipeline based on infrared images. It detects candidates, then tracks and re-identifies them using SIFT-features. In the depth modality, Barbosa et. al. [2] re-identifies by comparing various physical body measurements (anthropometrics) obtained from the depth image. Velardo and Dugelay [10] uses manually measured anthropometrics to prune the set of candidates for face recognition.

Finally, two papers combine several modalities. In [9] RGB is used for detection and re-identification, and depth for segmentation and pruning of re-id candidates. This is the same basic approach as in this work. In [8], thermal images and anthropometric measurements are added and the re-identification is performed in a truly multi-modal way with a combination of color histograms, SIFT features on thermal images, and anthropometric measurements obtained from depth images.

2.1 Datasets

Several public datasets exist, though mostly sets captured with traditional visible light sensors.

In other modalities, not many exist. For depth, the RGB-D Person Re-identification Dataset [2] is one option. It contains 79 people in 4 different scenarios: Walking slowly with outstretched arms, two instances of walking from a frontal viewpoint, and walking from a rear viewpoint.

For this work, we use our own dataset with a surveillance-like camera setup. We have three sequences: Novi, Basement, and Hallway. They all contain sequences of persons walking diagonally towards and past the sensor twice. Novi, which was also used in [9], contains 22 persons over 7800 frames (passes have varying lengths). Basement contains 35 persons over 7231 frames, and Hallway contains 10 persons over 4492 frames. Stats about the public as well as our own datasets can be seen in table H.1. The sequences were captured with Microsoft Kinect for Xbox. Example pictures from each sequence can be seen in fig. H.1.

Table H.1: Statistics on the three data sequences used in this work.

	Novi	Basement	Hallway
Number of persons	22	35	10
Number of frames	7800	7231	4492
Contains image sequences	Yes	Yes	Yes
Available modalities	RGB, depth	RGB, depth, thermal	RGB, depth, thermal

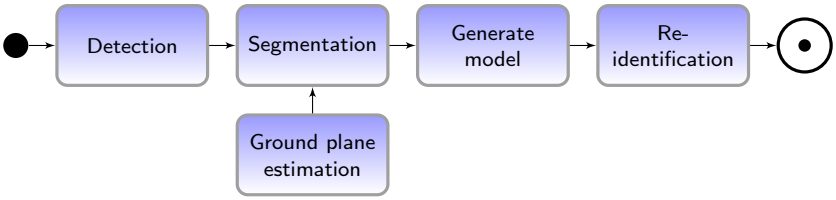


Fig. H.2: Illustration of the flow through the system.

3 Algorithm overview

This paper describes a full re-identification system which takes a raw RGB-D feed as input and outputs whether or not a passing person has been seen before, and if so, what the previous ID was. This is different from many other re-identification papers which most often describe a core algorithm without much focus on all the other system parts that must be in place to have an actual working system. The process requires several steps: Persons must be detected and segmented, they must be modeled, and finally re-identified. On top of the re-identification process comes the process of keeping tabs on the person database. A flowchart is shown in fig. H.2.

3.1 Detection and segmentation

The detection is done with a standard HOG-detector as first proposed by Dalal and Triggs [4]. The detector is trained on the INRIA Person Dataset introduced by the same paper. The detector runs on the RGB images and returns person bounding boxes.

The detected persons need to be segmented in further detail. The bounding box is not sufficient, since we do not want to capture features from the background. Segmentation is achieved with a flood fill in the depth image. Persons not crawling on the floor are conveniently separated from the background in the depth modality, so a

flood fill to similar pixels starting at the points

$$\mathbf{X} = \begin{bmatrix} 2/5 & 1/4 \\ 2/5 & 1/3 \\ 2/5 & 2/5 \\ 1/2 & 1/4 \\ 1/2 & 1/3 \\ 1/2 & 2/5 \\ 3/5 & 1/4 \\ 3/5 & 1/3 \\ 3/5 & 2/5 \end{bmatrix} \begin{bmatrix} b_w & 0 \\ 0 & b_h \end{bmatrix} + \begin{bmatrix} b_x & b_y \\ \vdots & \vdots \\ b_x & b_y \end{bmatrix}_{9x2} \tag{H.1}$$

where $\mathbf{X}$ is a 9x2 matrix containing the x and y coordinates of the flood fill points, b is the bounding box with subscript x, y, w, and h meaning top-left x-coordinate, top-left y-coordinate, width, and height respectively. The flood fill is performed at multiple positions to ensure that we have a stable object in the depth modality. A person is classified as stable if at least j depth points converge, i.e. the flood fill of these points fill out the same volume. For this implementation, $j = 4$.

Ground plane estimation

One problem with the flood fill is that at the feet of the subject, the fill is likely to spill onto the floor. To counter this, ground plane pixels on the depth image are removed. When the system is started initially, a ground plane is defined in the depth image. This is done by marking a number of points on the ground and performing a least squares solution of the bivariate polynomial:

$$z_{\text{poly}} = a_{00} + a_{01}x + a_{02}x^2 + a_{10}y + a_{20}y^2 + a_{11}xy \tag{H.2}$$

Although the floor is planar, the measurements of the floor from the Kinect depth sensor are representing the plane as a hyperbolic plane, thus stating the need for a bivariate polynomial. When the coefficients are determined, any pixel in the depth image close to the ground plane is colored black. Those pixels are the ones fulfilling the inequality in equation (H.3), where p is the pixel in question and t_{depth} defines the distance from the theoretical ground plane that is still considered part of that plane.

$$|z_{\text{poly}} - p_z| < t_{\text{depth}} \tag{H.3}$$

3.2 Person model

One of the objectives of this paper is to compare two types of multi-shot person models. They are both based on the two-part color histogram used in [9]: After a person is segmented, a color histogram is computed for the upper part of the body and the lower part of the body (as illustrated by the blue boxes in fig. H.3). Each color channel is divided into 20 bins, the individual channel histograms are concatenated, and finally the two part histograms are concatenated for a feature vector of $20 \cdot 3 \cdot 2 =$

Fig. H.3: The left image illustrates a detection. On the right, the person has been segmented in the depth image, and the blue boxes illustrates the boxes which are used as basis for the color histograms.

120 dimensions in the case of a 3 channel color space. In addition to the two modeling paradigms, 3 different color spaces were tested: RGB, HSV, and XYZ. For HSV and XYZ the luminance channels were removed to enhance lighting invariance, so in those cases the final histogram would be 80-dimensional and contain just the HS- and XZ-channels, respectively.

Two multi-shot schemes have been tested:

1) Mean histogram of all frames in a pass.

2) All frame-histograms saved individually.

In 1) the mean histogram is computed when a pass is over. Each bin is simply averaged:

$$m_i = \frac{1}{n} \sum_{j=0}^{n} h_{i,j} \text{ for } 0 \leq i < k \tag{H.4}$$

where m is the mean histogram, n is the number of frames in the pass, k is the number of bins in the histograms and $h_{i,j}$ is the value of bin i in histogram j.

In 2) no averaging takes place. Instead a pass is modeled after each histogram in it. See the following section on how each model is matched against the person database.

Both of the color-based models are augmented with a measure of the person's height. We use normalized height-to-border. This is the distance in pixels from the top of the person in the image, to the bottom of the frame, normalized by the depth of the observation. This reduces noise, as only one of the bounds of the height is now determined from the noisy depth sensor. It also allows for clipping.

In fig. H.4 height-to-border versus depth is plotted. Because the surface and field-of-view is the same for all who pass by the camera, the only change that will happen to the curve for people of different heights is a shift in its y-axis intercept. Instead of approximating the full curve, we go for the less computationally heavy option of modelling each pass with the mean of the depth-normalized height-to-border, designated γ, for all instances in the pass:

$$\gamma = \frac{1}{n} \sum_{i=0}^{n} g_i \cdot d_i \tag{H.5}$$

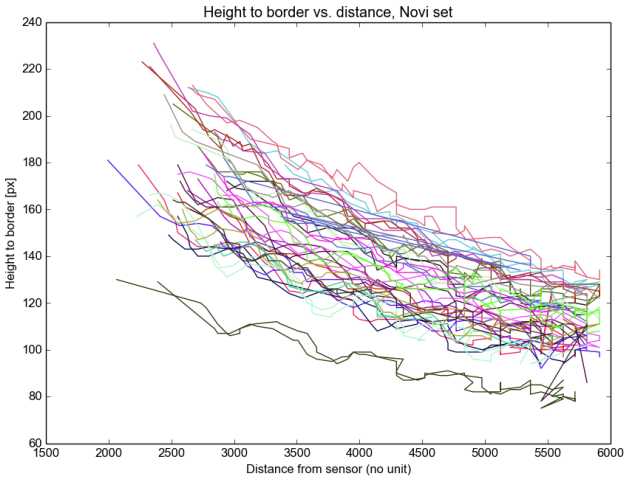


Fig. H.4: Curves depicting height-to-border versus distance for all tracks in a sequence. The curves are colored in pairs, such that two tracks of the same color are two passes by the same person. It can be seen that most lines are close to their partner of the same color, showing that the height measurement is stable across passes.

where g_i is the height-to-border for observation i in the pass, and d_i is the distance to the person in that observation. While the person is not completely flat, for the purpose of this normalization, we use the depth of the seed point described in equation H.1.

3.3 Re-identification

A pruning stage based on the height measurement is used before the re-identification. The height of the probe is compared to the gallery by means of the absolute difference in their heights. If the mean normalized height-to-border is more than t_h away from a candidate, the candidate is not considered a match for this subject. t_h is found from analyzing training data before running the system. The threshold t_h is set to the mean of the height difference between wrong matches in the training set.

When re-identifying, the model of the current pass is compared to those of the persons in the database, which is initially empty, but will be built as time progresses. Both the mean histogram and the histogram series model use the Bhattachariyya distance [3]:

$$d(H_1, H_2) = \sqrt{1 - \sum_I \frac{\sqrt{H_1(I)H_2(I)}}{\sqrt{\sum_I H_1(I) \cdot \sum_I H_2(I)}}} \tag{H.6}$$

where $d(H_1, H_2)$ is the distance between the histograms H_1 and H_2, and $H(I)$ is the value of bin I in the histogram H. The result is a number between 0 and 1, where 0 is a perfect match.

With mean histograms, where only two histograms - probe and gallery - are in-

volved, the distance itself is used, and the subject is either re-identified, ignored, or added to the database. With histogram series, the model comprise a series of histograms. In this case, each histogram in the probe model is compared to each histogram in the database. The probe then casts a vote for the ID of the gallery-model which contains the histogram it is closest to, if that is within a separately trained ignore threshold. The gallery-model with the most votes is selected as the best candidate, provided is has the majority (more than 50%) of the possible votes.

3.4 Mean histogram

The re-identification process is governed by two thresholds:

$$t_n\text{: New threshold: Subjects with } d(H_1, H_2) > t_n \text{ are added as new persons} \tag{H.7}$$

$$t_i\text{: Ignore threshold: Subjects with } d(H_1, H_2) <= t_i \text{ are re-identified} \tag{H.8}$$

This implicates that subjects with $t_i < d(H_1, H_2) <= t_n$ are ignored, because they are too similar to other subjects, without being similar enough to trust the identification.

The thresholds are learned beforehand by observing a training set. The distances between all mean histograms in the training set are computed and stored in the set $\mathcal{D}$ and divided into two sets $\mathcal{D}^c$ and $\mathcal{D}^w$ where $\mathcal{D}^c$ contains distances between different observations of the same person and $\mathcal{D}^w$ contains distances between histograms of different persons:

$$\mathcal{D}^c = \{\mathcal{D} | id(H_1) = id(H_2) \text{ in } d(H_1, H_2)\} \tag{H.9}$$

$$\mathcal{D}^w = \{\mathcal{D} | id(H_1) \neq id(H_2) \text{ in } d(H_1, H_2)\} \tag{H.10}$$

where $id(\bullet)$ is the person id connected with a histogram. The thresholds are then computed as:

$$t_n = \overline{\mathcal{D}^w} - 2 \cdot \sigma(\mathcal{D}^w) \tag{H.11}$$

$$t_i = \overline{\mathcal{D}^c} + \sigma(\mathcal{D}^c) \tag{H.12}$$

where $\overline{\bullet}$ denotes mean and $\sigma(\bullet)$ denotes standard deviation.

3.5 Histogram series

The re-identification for the histogram series model uses many of the same principles of the mean histogram model, but is adapted to use many more histograms for each subject to encompass variations in lighting and pose. A histogram is computed for each frame in the pass of a subject and they are then compared to all histograms already in the database. When the shortest distance d_s to any gallery-histogram is less than t_i, the associated person id, p_s receives a vote. Thus, each subject histogram contributes with up to 1 vote, for a theoretical total of $len(\mathbf{H})$ votes: the number of histograms in the current pass. If there are no histograms in the pass, the subject

is ignored. If any person in the gallery has received more than half the theoretical maximum, the subject is re-identified as him. If no gallery person satisfies this requirement, the subject is added as a new person.

It is worth noting that this method has no explicit option of ignoring the subject in case it is uncertain, other than in the case where no histograms exist.

4 Evaluation

6 permutations of the system have been tested on 3 different sequences (see section 2.1). The 2 different multi-shot models have both been tested in 3 different color spaces: RGB, HSV, and XYZ.HSV and XYZ have been tested since they both model color closer to how the human eye sees it, and more specifically because they allow for exclusion of the luminance so that differing lighting conditions should affect performance less. That means that for the following tests all three RGB channels were used, in the HSV case only HS were used, and with XYZ only XZ were used.

The performance of the system varies with the order the persons are passing by the camera. If a person that is very hard to re-identify passes by the camera in the first two passes without any other entries in the database, odds are that he will be correctly re-identified. However, if a similar person enters the database before the second pass of person 1, they might be confused with each other and thus lower the performance. To even out this effect, all results presented below are averages of 100 runs where the subjects enters the system in random order. That should sufficiently even out any "lucky" or "unlucky" orderings and provide accurate results. For each run, all thresholds have been trained on a random subset of 20% of the sequence, which is then excluded from the rest of the run. The effect of the training set selection should also average out.

The re-identification performance can be characterized with 5 parameters:

1. Correct new
2. Wrong new
3. Correct ID
4. Wrong ID
5. Ignored

The first two describes how well the system distinguishes between known persons and new persons. Ideally, there should be no wrong new, as they are persons that are already in the database and should have been re-identified. Correct ID and wrong ID comprises the subjects that are neither ignored, correct new, nor wrong new, but are re-identified. Finally, ignored are the ones that are not handled because they are neither close enough to an existing person to be re-identified, nor different enough from the existing persons to be added to the database.

The results of the tests can be seen in table H.3. Sequence length and detection performance varies greatly between sequences, as seen in table H.2. Note that the Hallway sequence contains many shorter tracks, meaning that generalization, as well as the benefit from the multi-shot approach, declines heavily.

Table H.2: Statistics on the amount of observations of captured persons for each sequence. The numbers are based on the amount of times a single person was detected and modeled in a single pass.

	Basement seq.	Hallway seq.	Novi seq.
Mean observation length:	25.5	10.3	40.7
Median observation length:	24	11.5	41
Minimum observation length:	4	2	5
Maximum observation length:	38	25	57

Generally, the mean histogram and histogram series approaches perform equally when looking at the percentage rates of the identification. The differences between the two approaches are most profound in the Basement and Novi sequences. The histogram series approach contains no ignore category which leads to a higher number of wrong new identifications than compared with the mean histograms. However, the method returns a significantly lower number of wrong identifications in both sequences. It is seen from the standard deviation of that the mean histogram exhibits a more stable performance than the histogram series on correct identifications whereas the opposite seems to be the case for wrong identifications. The number of wrong identifications is low across the board, so the weak spots are the wrong new- and ignored-counts which are rather high. Most new passes are correctly classified as such, at around 29-32 of 35 in the basement sequence, 8/10 and 21/22 in the Hallway and Novi sequences respectively.

Table H.3: Re-identification performance of the 6 system configurations on 3 different sequences. All numbers are averaged over 100 runs with random enrollment order. The standard deviation of the results are shown in parenthesis.

Basement sequence

		Correct new	Wrong new	Correct ID	Wrong ID	Ignored	% correct	% wrong
RGB	Mean histogram	29.31 (2.92)	4.53 (7.10)	11.76 (5.34)	1.22 (2.50)	8.18 (6.82)	90.62 %	9.38 %
	Histogram series	32.61 (1.44)	8.64 (7.37)	13.00 (7.07)	0.74 (2.14)	**0.00** (0.00)	94.60 %	5.40 %
HS	Mean histogram	28.75 (3.23)	**3.20** (7.36)	12.57 (5.81)	1.18 (2.83)	9.30 (6.78)	91.43 %	8.57 %
	Histogram series	**32.71** (1.37)	8.13 (7.77)	**13.49** (7.51)	**0.67** (2.27)	**0.00** (0.00)	**95.24** %	**4.76** %
XY	Mean histogram	29.02 (3.26)	4.72 (7.12)	11.24 (5.15)	1.68 (3.01)	8.34 (7.35)	86.97 %	13.03 %
	Histogram series	32.54 (1.45)	8.88 (7.34)	12.59 (6.87)	0.98 (2.33)	**0.00** (0.00)	92.78 %	7.22 %

Hallway sequence

		Correct new	Wrong new	Correct ID	Wrong ID	Ignored	% correct	% wrong
RGB	Mean histogram	8.36 (1.00)	4.07 (2.34)	1.15 (1.37)	0.61 (1.01)	0.81 (1.62)	65.34 %	34.66 %
	Histogram series	8.59 (0.77)	4.45 (2.10)	1.30 (1.57)	0.66 (1.18)	**0.00** (0.00)	66.33 %	33.67 %
HS	Mean histogram	8.15 (1.17)	**3.95** (2.41)	1.34 (1.61)	**0.39** (0.78)	1.17 (2.13)	**77.46** %	**22.54** %
	Histogram series	**8.61** (0.62)	4.44 (2.14)	**1.44** (1.72)	0.51 (0.76)	**0.00** (0.00)	73.85 %	26.15 %
XY	Mean histogram	8.37 (0.96)	4.11 (2.29)	1.15 (1.37)	0.63 (1.10)	0.74 (1.58)	64.61 %	35.39 %
	Histogram series	8.57 (0.71)	4.33 (2.14)	1.37 (1.58)	0.73 (1.22)	**0.00** (0.00)	65.24 %	34.76 %

Novi sequence

		Correct new	Wrong new	Correct ID	Wrong ID	Ignored	% correct	% wrong
RGB	Mean histogram	21.15 (1.40)	3.79 (2.98)	9.68 (2.73)	0.25 (1.31)	1.13 (1.89)	97.48 %	2.52 %
	Histogram series	**21.51** (0.70)	9.12 (4.21)	5.31 (4.24)	**0.06** (0.42)	**0.00** (0.00)	**98.88** %	**1.12** %
HS	Mean histogram	20.64 (1.86)	**2.42** (2.13)	**10.52** (2.38)	0.44 (1.45)	1.98 (3.21)	95.99 %	4.01 %
	Histogram series	21.48 (0.77)	9.18 (4.53)	5.23 (4.59)	0.11 (0.91)	**0.00** (0.00)	97.94 %	2.06 %
XY	Mean histogram	21.14 (1.17)	4.89 (3.48)	8.83 (3.10)	0.34 (1.26)	0.80 (1.74)	96.29 %	3.71 %
	Histogram series	21.46 (0.87)	10.12 (3.91)	4.31 (3.93)	0.11 (0.91)	**0.00** (0.00)	97.51 %	2.49 %

Table H.4: Comparison of re-identification performance with and without the height-based candidate pruning step.

		Without height		**With height**		**Difference**	
		% correct	% wrong	% correct	% wrong	% correct	% wrong
Basement	Mean hist.	82.17 %	17.83 %	90.67 %	9.33 %	8.50 %	-8.50 %
	Hist series	87.28 %	12.72 %	94.21 %	5.79 %	6.93 %	-6.93 %
Hallway	Mean hist.	64.64 %	35.36 %	69.14 %	30.86 %	4.50 %	-4.50 %
	Hist. series	67.34 %	32.66 %	68.47 %	31.53 %	1.10 %	-1.10 %
Novi	Mean hist.	92.03 %	7.97 %	96.59 %	3.41 %	4.56 %	-4.56 %
	Hist. series	96.50 %	3.51 %	98.11 %	1.89 %	1.61 %	-1.61 %
Average		81,66 %	18,34 %	86,20 %	13,80 %	**4,53 %**	-4.53 %

The benefit of the ignore-functionality in the mean histogram model is illustrated in fig. H.5. Blue columns are a histogram of distances between mean histograms of the same person, while red columns are a histogram of distances between different persons. The overlap between these shows that is it not possible to achieve perfect classification with a 1d decision boundary in this case. To counter this, an ignore zone is introduced - the space between the green and the yellow line, the thresholds, which can to some extent mitigate the effects of this overlap. In reality, when training on a subset of the data, the ignore zones are generally wider than in this example. It is possible that a classification in a higher dimensional space would work better and allow discarding the ignore zone.

Table H.4 shows how the height-based pruning step improves the re-id rates across all methods. By discarding obviously wrong candidates based on height, the correct re-id rate goes up by 4.53 percentage points on average.

We have been unable to compare our results to the work of others, as they do not present full-flow systems, but rely on tightly pre-cropped images of persons. Furthermore, our system needs depth images as well as RGB, so no existing dataset has been compatible. We also do not present CMC-curves as that ranking system works poorly for on-the-fly enrollment systems, where, in many cases, there are simply not enough entries in the database to do a proper ranking.

We can, however, compare some of our results to the work previously presented in [9]. Not all stats are directly comparable, but the correct and wrong ID rates are. In that work, they are 68% and 0%, with an ignore rate of 24%. The system presented here has a much higher correct ID rate, but at the cost of a somewhat higher wrong ID rate.

5 Conclusion

This work presented a re-identification system using RGB-D data and compared several model and color space configurations. It introduces 3 new, different re-identification sequences for testing, and goes through all stages from candidate detection to identification. Furthermore, it investigates how to handle online enrollment of subjects, a subject few previous works have touched. Future work includes more

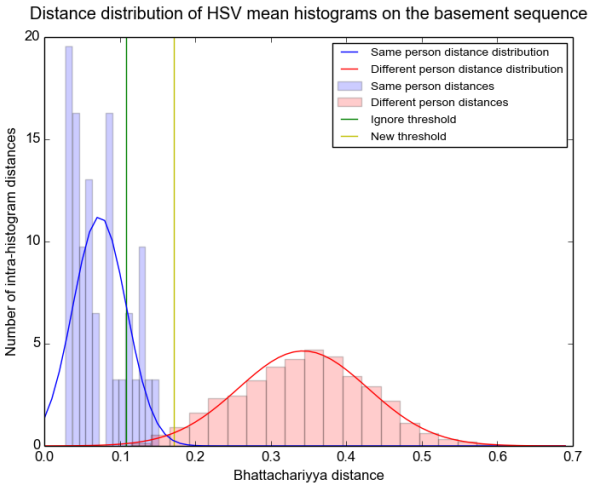


Fig. H.5: Distribution of distances between histograms in the full basement sequence. There is a clear overlap of distances between histograms from the same person and histograms from different persons. When using a distance threshold to classify, this will result in wrong identifications. The ignore-threshold allows to remove the distances that are the most affected by this overlap.

sophisticated multi-shot models, and enhancing the system to cope with multiple, co-occluding subjects in crowded environments.

Bibliography

[1] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. "Learning to Match Appearances by Correlations in a Covariance Metric Space". In: *ECCV (3)*. Vol. 7574. LNCS. Springer, 2012, pp. 806–820. ISBN: 978-3-642-33711-6.

[2] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. "Re-identification with RGB-D Sensors." In: *ECCV Workshops (1)*. Vol. 7583. LNCS. Springer, 2012, pp. 433–442. ISBN: 978-3-642-33862-5.

[3] G. Bradski and A. Kaehler. "Learning OpenCV". In: O'Reilly, 2008. Chap. 7, pp. 201–202.

[4] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *CVPR*. 2005.

[5] M. Demirkus, K. Garg, and S. Guler. "Automated person categorization for video surveillance using soft biometrics". In: *Biometric Technology for Human Identification VII*. 2010.

[6] Gianfranco Doretto, Thomas Sebastian, Peter H. Tu, and Jens Rittscher. "Appearance-based person reidentification in camera networks: problem overview and current approaches". In: *J. Ambient Intelligence and Humanized Computing* 2.2 (2011), pp. 127–151.

[7] Kai Jüngling and Michael Arens. "Local Feature Based Person Reidentification in Infrared Image Sequences." In: *AVSS*. IEEE Computer Society, 2010, pp. 448–455. ISBN: 978-0-7695-4264-5.

[8] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund, and Sergio Escalera. "Tri-modal Person Re-identification with RGB, Depth and Thermal Features". In: *9th Workshop on Perception Beyond the Visible Spectrum, CVPR Workshops*. IEEE, 2013, pp. 301–307.

[9] Andreas Møgelmose, Thomas B. Moeslund, and Kamal Nasrollahi. "Multimodal Person Re-Identification using RGB-D Sensors and a Transient Identification Database". In: *International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2013, pp. 1–4.

[10] C. Velardo and J. Dugelay. “Improving Identification by Pruning: A Case Study on Face Recognition and Body Soft Biometric”. In: *WIAMIS*. IEEE, 2012, pp. 1–4. ISBN: 978-1-4673-0791-8.

[11] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. “Unsupervised Salience Learning for Person Re-identification”. In: CVPR. 2013.

[12] W. Zheng, S. Gong, and T. Xiang. “Person re-identification by probabilistic relative distance comparison”. In: *CVPR*. IEEE, 2011, pp. 649–656.

Paper I

Tri-modal Person Re-identification with RGB, Depth and Thermal Features

Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund, and Sergio Escalera

The paper has been published in the
Proceedings of the 9th Workshop on Perception Be- yond the Visible Spectrum, CVPR Workshops, pp. 301–307, 2013.

The layout has been revised.

Abstract

Person re-identification is about recognizing people who have passed by a sensor earlier. Previous work is mainly based on RGB data, but in this work we for the first time present a system where we combine RGB, depth, and thermal data for re-identification purposes. First, from each of the three modalities, we obtain some particular features: from RGB data, we model color information from different regions of the body; from depth data, we compute different soft body biometrics; and from thermal data, we extract local structural information. Then, the three information types are combined in a joined classifier. The tri-modal system is evaluated on a new RGB-D-T dataset, showing successful results in re-identification scenarios.

1 Introduction

Person re-identification is about recognizing people who have passed by a sensor earlier. It is useful in many places where it is desirable to obtain knowledge of the flow of people: airports, transit centers, shopping malls, amusement parks, etc. It can either be knowledge of a single person's movement, or movement patterns in general by combining the patterns of many people. In some cases it is possible to set up a system, which is able to view the entire scene, as in [20, 17]. However, in indoor scenes it is often not feasible to place one camera with a full overview. This is where re-identification enters play. It allows the system designer to place sensors at certain bottlenecks and identify people when they pass these.

Re-identification has the specific distinction from e.g. biometric access control systems that it must be able to enroll new people on-the-fly and without their specific collaboration. On the other hand, the recognition performance does not necessarily have to be as strong as in access control systems, since re-identification systems are more concerned with the general trend of movement as opposed to the movement of each individual.

Re-identification has been an active research area for the past decade, but almost exclusively focused on standard RGB-data. This makes sense since many venues have a large network of already installed RGB surveillance cameras. However, as new and more advanced sensor types become cheaply available, we believe it is time to extend the work to multiple modalities. This is the exact focus of this work, where we present a novel approach that integrates RGB, depth, and thermal data in a re-identification system. An example of RGB, depth, and thermal images for a subject in our dataset is shown in Figure I.5.

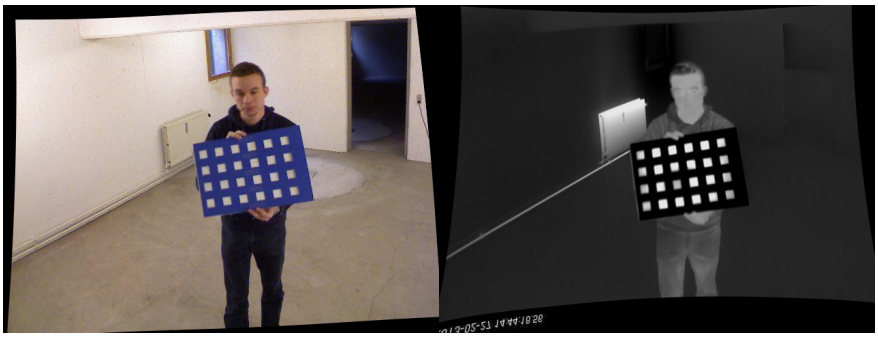
This paper is structured as follows: Section 2 briefly covers the existing work done on the topic of re-identification, with special focus on the few multi-modal and/or non-RGB-based contributions. Section 3 describes how the inputs from the three modalities are aligned. In sections 4 and 5, the features and re-identification methods are presented. Section 6 shows the dataset and covers the results our system achieves on it. Finally, section 7 concludes the paper.

2 Related work

In [6] soft-biometrics based on RGB data are used to track people across different cameras. Both body and facial soft biometrics are extracted and combined in the final system. The body soft biometrics are all related to color: hair, skin, upper, and lower body clothing. In [7] the notion of tracking people across a multi-camera setup is also followed. Different soft biometric features are reviewed and discussed in the context of re-identification. A part-based appearance approach is found to perform the best, but being sensitive to how the object is divided into parts. In [8] each person is also divided into parts from which features are extracted. The division is here based on finding symmetry axes and the soft biometric features are color histograms, stable color regions and highly structured patches that reoccur. A division is also applied in [10] using similar features. A boosting approach is then introduced to select the most discriminative features. In [1] a similar idea is proposed, i.e., a more reliable classification can be obtained if only the most discriminative features are used for each image region. Moreover they model the uncertainties (covariances) of each feature to improve their results. In [23] a person is divided into six horizontal stripes where each is described in terms of color and texture. The novelty of the work if the formulation of the re-identification problem as a matter of learning the optimal distance measure that minimizes the probability of miss-classification.

All of the above approaches are based on RGB data. Using multi-modal sensing in re-identification is a very new concept and so far only a few works have been reported. In [21] a two-stage recognition approach is followed. First soft-biometrics based on depth data are extracted and secondly RGB data are used in the final classification step. The depth-based soft biometrics are anthropometric measurements and estimated manually. The key finding is that soft biometrics can be used as a pruning step in a recognition system. While this is very interesting, the introduction of manual measurements is not desirable for an automatic re-identification system. In [2] a re-identification method based solely on depth features is presented. The work uses several normalized measures of body parts, calculated from joint positions. Measures of the body's "roundness", which roughly estimates the volume of the torso, are included. High depth resolution is required for this to work and hence it is only suitable when subjects are close to the sensor. The paper is focused solely on the re-identification step and does not treat identification or extraction of joints. In [12] thermal data are used in a re-identification system. The work expands the work reported in [13] where SIFT features are used to model each person. They work on gait data from a side view and can thou track each body part reliably. From each of these a codebook signature is learned over time and combined with the spatial feature distribution found using an Implicit Shape Model.

As opposed to the works described above, in this paper we introduce a truly multi-modal approach based on RGB, depth and thermal data. Moreover, our system is fully automated both in terms of feature extraction, but also when it comes to enrollment.



(a) **(b)**

Fig. I.1: Stereo rectified multimodal imagery in the (a) RGB and (b) thermal domains.

3 Registration

Since no sensor is able to capture all three modalities at once, a registration of the inputs must take place allowing to map a specific point from one modality to the others. In this work, the Microsoft® Kinect™ for XBOX360 has been used to capture RGB and depth data. A thermal camera (AXIS Q1922) was mounted straight over the Kinect's RGB camera lens with a distance between the lens centers of 70 mm. For registering the tri-modal imagery of this work, we need only to register images from the thermal and visual modalities, as the Kinect provides a factory calibrated registration between the RGB and depth data.

Traditional image registration techniques used for spatially aligning stereo imagery cannot be directly applied to the thermal-visible domain due to the fundamental physical differences of the two modalities, thus rendering the process of finding corresponding features in both imagery is unfeasible. In our setup, objects appear at distances between 1 and 4 meters from the cameras, which makes methods like infinite homography and stereo geometric unusable [15]. Instead we first use stereo rectification to transform the epipolar lines to lines parallel with either the x or y axis [9]. This reduces the search for corresponding points to one dimension. Next we apply the notion that the distance between corresponding points in the two images is inversely proportional to the depth of the points if the cameras are only translated with respect to each other [9]. Since the epipolar lines are transformed to lie along the image scanlines, the disparity between corresponding points will lie mainly either on the x or y axes, and we may thus find the relationship between the inverted depth and the induced disparity and use this property for rectifying the images.

The stereo calibration requires the knowledge of the intrinsic and extrinsic camera parameters of both cameras. In order to determine these, we use the calibration board proposed by [22] with an A3-sized cut-out checkerboard and a heated plate as a viable backdrop. By using standard camera calibration and stereo geometric tools we are able to rectify both images as seen in Figure I.1.

We used 34 image pairs of the calibration board distributed throughout the entire

scene for the calibration of the cameras. For each corner of the chessboard in each image, we extract the corresponding depth. The configuration of cameras placed vertically implies that the disparity of the points in the rectified image lies mainly on the x-axis. Therefore, we use a robust curve fitting tool to find a linear regression that fits the disparity in the x-direction as a function of the inverted distance in the z-direction. The regression is computed off-line for all calibration points and stored for online lookup of the displacement. The result of this procedure is a direct pixel-to-pixel correspondence between the different images.

4 Multi-modal features

The proposed system uses a combination of RGB, depth, and thermal features to perform the re-identification task. This section explains how the feature extraction is performed for each modality. Before the extraction, the subject must first be located at pixel level. The foreground segmentation of the subject is performed on the depth image by means of Random Forest [19]. This process is performed computing random offsets of depth features as follows:

$$f_\theta(\mathcal{D}, \mathbf{x}) = \mathcal{D}_{\left(\mathbf{x}+\frac{\mathbf{u}}{\mathcal{D}_\mathbf{x}}\right)} - \mathcal{D}_{\left(\mathbf{x}+\frac{\mathbf{v}}{\mathcal{D}_\mathbf{x}}\right)}, \tag{I.1}$$

where $\theta = (\mathbf{u}, \mathbf{v})$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\mathbf{2}$ is a pair of offsets, depth invariant. Thus, each θ determines two new pixels relative to $\mathbf{x}$, the depth difference of which accounts for the value of $f_\theta(\mathcal{D}, \mathbf{x})$. Using this set of random depth features, Random Forest is trained for a set of trees, where each tree consists of split and leaf nodes (the root is also a split node). Finally, a final pixel probability of body part membership l_i is obtained as follows:

$$P\left(l_i|\mathcal{D}, \mathbf{x}\right) = \frac{1}{\tau}\sum_{j=1}^{\tau} P_j\left(l_i|\mathcal{D}, \mathbf{x}\right), \tag{I.2}$$

where $P\left(l_i|\mathcal{D}, \mathbf{x}\right)$ is the PDF stored at the leaf, reached by the pixel for classification $(\mathcal{D}, \mathbf{x})$ and traced through the tree j, $j \in \tau$. After this process, the foreground segmentation mask of the subject is transformed to the coordinate system in the two other modalities, and the features are extracted.

The system uses multi-shot person models. Thus, a person is not modeled based on only one frame, but on all frames in a pass. A pass is defined as the act of entering the frame, walking by the camera, and exiting it. In our dataset only one person is present at a time, so no tracking is necessary. Next, we describe how the features from each modality are described and fused in order to perform the on-line re-identification task. Figure I.2 summarizes the main modules, modalities and strategies considered in the proposed re-identification system.

4.1 RGB features

After foreground segmentation is performed, the features that are used for the RGB modality are color histograms in two parts, as shown in Figure I.3(a). One histogram H_U^{RGB} is derived from the upper body, one H_L^{RGB} from the lower. This is done for each frame in which the subject is detected. A histogram of 20 bins is created for each

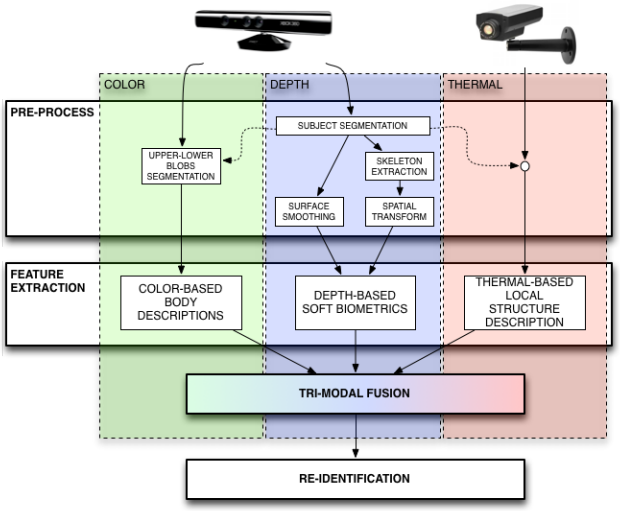
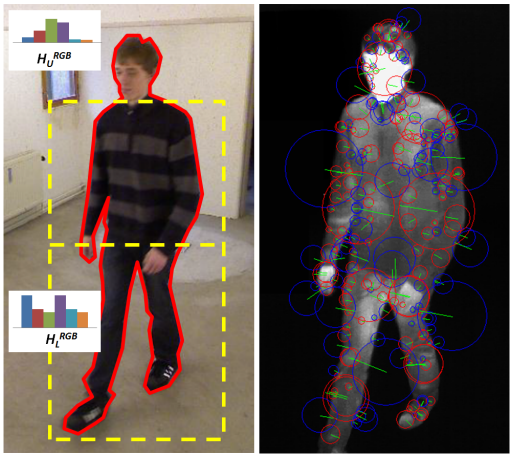


Fig. I.2: Pipeline of the proposed tri-modal re-identification system.



(a) (b)

Fig. I.3: (a) Histograms of RGB color distributions for upper body H_U^{RGB} and lower body H_L^{RGB} parts of the subject. (b) Detected SURF keypoints on the thermal modality.

channel, for a total of 60 bins per body part. Thus, in total the RGB feature vector has 120 dimensions, and one is created per frame. After a pass ends, the histograms are averaged, and the final feature vector is the mean across the frames.

4.2 Depth features

Given an input depth frame containing a subject (Figure I.4(a)), and once the pixel-ground segmentation of the subject into body parts is performed, the skeleton is also extracted applying Mean Shift [19] (Figure I.4(b)). Since our dataset contains only raw images, the built-in skeleton-extraction from the Kinect could not be used. Then, the subject point cloud is spatially transformed in order to align the skeleton with the camera frame coordinate system by means of an affine three-dimensional transformation of the point cloud (Figure I.4(c)). Note that because of the 3D transformation we loose some information of the body surface due to the lack of information inherent to the viewpoint. Thus, the noisy subject's surface is smoothed (Moving Least Squares surface reconstruction method) and up-sampled to fill the holes (Figure I.4(d)). Now we can compute soft biometrics from the corrected 3D skeleton and the 3D surface of the aligned body, which can be then inversely transformed to return to the original space and estimate real measurements of the body. From a given depth frame $\mathcal{D}_i$, information invariant to the rotation of the subject with respect to the camera viewpoint can now be extracted. In particular, we have estimated three sets of soft biometrics:

Frontal curve model: The model encodes the distances from the points in subject's surface (transformed and smoothed, as seen in Figure I.4(d)) to their corresponding projection line, either head-to-neck or neck-to-torso line. These distances in millimeters are encoded in a real-valued vector $\mathbf{f}_i$, resampled to size 150 and equalized for normalization purposes (Figure I.4(e)).

Thoracic geodesic distances: Corresponds to the vector $\mathbf{g}_i$. It contains the length of lines on the body surface from one side of the body to the other. The area in which these are found is the trapezoid defined by left shoulder, right shoulder, right hip, and left hip, and each entry of $\mathbf{g}_i$ contains the geodesic distance in millimeters of a horizontal line in the trapezoid projected to the surface of the torso. $\mathbf{g}_i$ is resampled to size 90 (Figure I.4(f)).

Anthropometric relations: Given the extracted body skeleton, the lengths of 7 inter-joint segments connecting the body parts, as shown in Figure I.4(c), are computed and stored as $\mathbf{a}_i$.

Thus, the vector representing the set of depth features for a subject in the scene at a particular depth frame $\mathcal{D}_i$ is defined as:

$$\delta_i = \{\mathbf{f}_i, \mathbf{g}_i, \mathbf{a}_i\},$$

where $|\delta_i| = 247$. Finally, the vector describing the subject pass $D = \{F, G, A\}$ is computed by averaging the set of the standardized frame-level depth feature vectors $\{\delta_1, ..., \delta_N\}$ as:

$$D = \frac{1}{N} \sum_{j \in N} \frac{\delta_j - \bar{\delta}}{\sigma_\delta}, \tag{I.3}$$

where $|D| = 247$, and $\bar{\delta}$ and σ_δ correspond to the mean depth vector and the vector of the standard deviations, respectively. Moreover, as a previous step to this

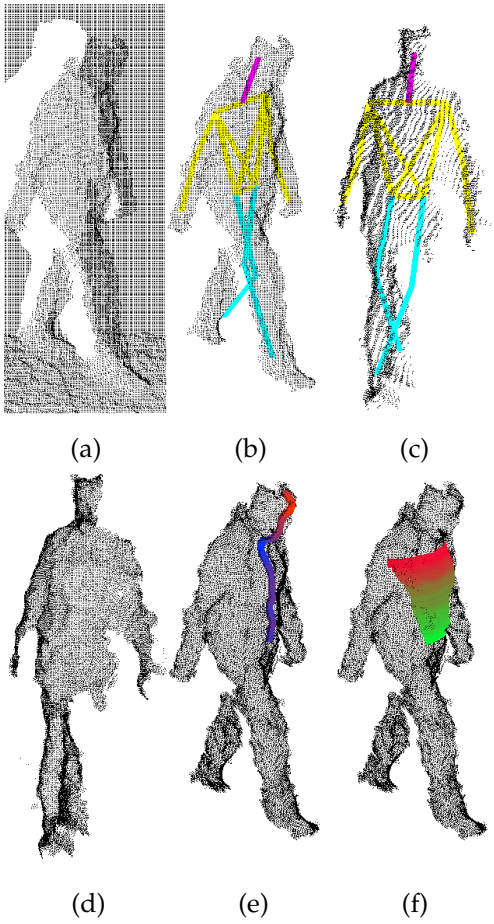


Fig. I.4: (a) The raw depth data. (b) The pixel-ground segmentation of the subject and the skeleton. (c) After aligning the skeleton with the camera frame. (d) Smoothed data. (e) Vertical projection lines. (f) Geodesic distances.

computation and due to the noisy nature of the captured depth data (clothes deformation, waving arms in front of the torso, and so forth), the possible outliers are detected and discarded in each δ_i. This step consists also in standardizing the set of depth feature vectors but to a modified Z-score [11] and discarding those values higher than 3.5 in absolute value.

4.3 Thermal features

Since the thermal images contain no color information, the color histogram approach does not work here. Instead, SURF[3] is employed. Within the contour supplied by the detection stage, SURF-descriptors are extracted. There is no fixed number of descriptors, all that are above a certain quality threshold are extracted. A typical number

is around 150 descriptors per subject per frame, depending on the contour's size and quality. As opposed to the RGB histograms there is no direct way to average the descriptors, so the model for people in the thermal modality is all SURF descriptors of the subject extracted over all frames in a pass. We define the set of detected and described SURF points as S, see Figure I.3(b).

5 Re-identification

In order to perform the re-identification task, previously computed feature vectors for the three modalities have to be fused and analyzed to classify each subject. The process has two steps:

1. Determine whether the subject is a new or an already known person.
2. Do one of the following two tasks:
   (a) If known, determine the ID of the person.
   (b) If new, enroll the person.

In step 1, a comparison of the current subject with the list of known subjects is done. Taking into account that the set of known persons is built on-the-fly, for the first evaluations only a few comparisons have to be performed.

To estimate whether the subject has to be considered new or re-identified, we compute the following confidence score based on the combination of the three modalities scores:

$$C(U_1, U_2) = \alpha \cdot d_{\text{RGB}}(H_1, H_2) + \beta \cdot \frac{1}{d_{\text{depth}}(D_1, D_2)} + \\ + \gamma \cdot \frac{1}{d_{\text{thermal}}(S_1, S_2)},$$

where $U_1 = \{H_1, D_1, S_1\}$ is the set of three modality descriptors (H_1 color histograms, D_1 depth feature vectors, and S_1 SURF descriptors on the thermal data) for a user in the dataset, and $U_2 = \{H_2, D_2, S_2\}$ are the three sets of descriptors for a new test subject. Coefficients α, β, and γ assigns a proper weight to each of the three modalities scores in a late fusion fashion so that $\alpha + \beta + \gamma = 1$. The weights are static and were set based on experimentation, but for future work, and especially larger datasets, a learning approach for the weights would have to be investigated. The higher the output of $C(U_1, U_2)$, the more reliable re-identification. Because $d_{\text{depth}}(D_1, D_2)$ and $d_{\text{thermal}}(S_1, S_2)$ returns low values in case of good identifications, the reciprocal is used when fused.

For comparing two subjects in the RGB-modality, the Bhattacharyya distance [5] is used:

$$d_{\text{RGB}}(H_1, H_2) = \sqrt{1 - \sum_I \frac{\sqrt{H_1(I) H_2(I)}}{\sqrt{\sum_I H_1(I) \cdot \sum_I H_2(I)}}}, \tag{I.4}$$

where $d_{\text{RGB}}(H_1, H_2)$ describes the distance between histograms H_1 and H_2, and $H(I)$ is the value of bin I in the histogram H. The distance is a number between 0 and 1, where 0 is a perfect match.

For comparing across subjects in the depth modality $D = \{F, G, A\}$, the following similarity measure is computed:

$$d_{\text{depth}}(D_1, D_2) = W_F(1 - \exp^{-\sum_i w_i (F_1^i - F_2^i)^2}) +$$

$$+W_G(1-\exp^{-\sum_j w_j(G_1^j-G_2^j)^2})+ \\ +W_A(1-\exp^{-\sum_k w_k(A_1^k-A_2^k)^2}). \quad \text{(I.5)}$$

One distance is computed for each of the three depth features, which is in the range [0..1], the lower the distance, the higher the similarity. Coefficients W_F, W_G, and W_A assigns a proper weight to each of the three types of depth feature sets so that $W_F + W_G + W_A = 1$. Moreover, individual feature weights w assign a weight to each particular depth feature value, pre-computed based on a training stage applying ReliefF [18]. In out case the variables were set to $W_F = 0.8$, $W_G = 0.1$, and $W_A = 0.1$.

In the thermal domain, the SURF-descriptors are matched against each other with no spatial information resolved. Each matched feature contributes a vote. Thus the metric is the number of votes for a specific known person across all the frames in the model:

$$d_{\text{thermal}}(S_1, S_2) = \sum_{N_{S_2}} H(n_{\text{votes}}(S_1, S_2)), \quad \text{(I.6)}$$

where $n_{\text{votes}}(S_1, S_2)$ computes the number of matches between SURF descriptors S_1 on the reference image and SURF descriptors S_2 on the test image based on Euclidean distance criterion. H refers to the Heaviside step function, ensuring that each frame in a pass can only contribute one vote, and N are the frames in the model for S_2.

5.1 Determine if new

In order to determine if a person is new, once values for α, β, and γ are established based on a cross-validation of a training stage, two thresholds, T_N and T_R are also experimentally computed. If $C < T_N$, the subject is considered new. If $C > T_R$ the subject is assigned a known ID (re-identified). Since a false positive is more serious than a false negative in re-identification, we have a buffer zone when $T_N \leq C \leq T_R$ where the system ignores the subject because we are uncertain whether it is a new person or just a bad match to an existing one. In our system we used $T_N = 6$ and $T_R = 10$, but the exact value of the thresholds seemed to be relatively flexible.

5.2 ID determination

The assignment of an ID to an already existing user for re-identification is straightforward using the confidence score C obtained from the previous step. If the user has been determined as already known, it means that the majority of votes are given to a particular user ID which is the one assigned in the re-identification task.

6 Evaluation

Several re-identification datasets with RGB [10, 16] and RGB-D data [2] exist, but to the best of our knowledge no dataset containing all three modalities exists. We have therefore recorded a novel re-identification tri-modal dataset.

The dataset consists of 35 people passing by the sensors twice for 70 passes in total. The vantage point is up and slightly off to the side to mimic a classic surveillance

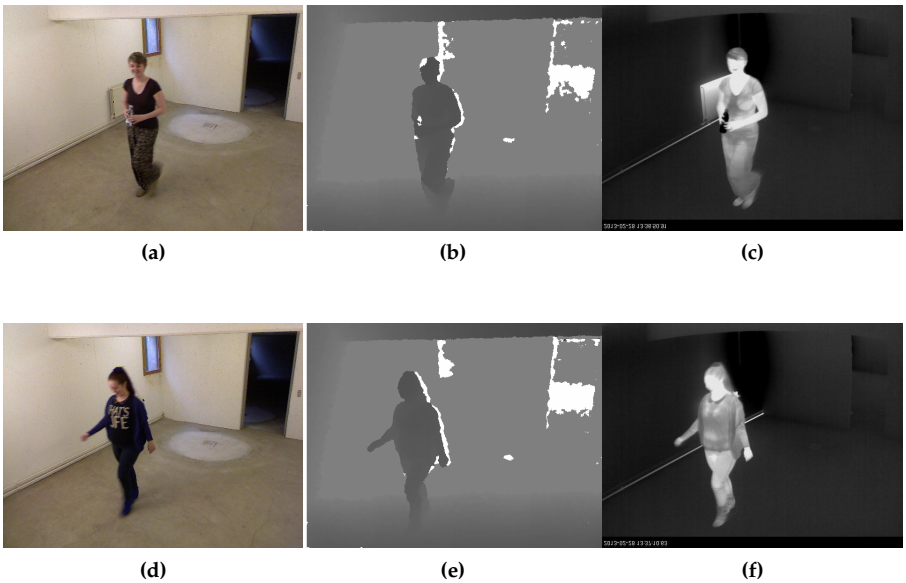


Fig. I.5: Sample images from the tri-modal dataset. left, middle, and right are RGB, depth, and thermal, respectively.

camera setup. All images are 640×480 pixel. Some sample images from each modality are shown in Figure I.5.

The tests were conducted by first extracting the aforementioned features from all passes. As this system is a re-identification system with online enrollment, there is no explicit training phase. Instead, the persons are enrolled if they are very different from previous seen persons.

Since the order of passing will influence the re-identification performance, the system was tested in a random 5-cross validation. We tried the different sets of modalities as input features and found that the best combination of features is the late fusion considering the three sets of modalities with weights: $\alpha = \frac{1}{3}$, $\beta = \frac{1}{3}$, and $\gamma = \frac{1}{3}$ to fit the tri-modal scheme. The results are presented both individually and averaged in terms of: A) passes correctly classified as a new person, B) passes wrongly classified as a new person, C) the number of correctly re-identified persons, D) the number of wrongly re-identified persons, and E) the number of persons ignored, see Table I.1.

If an application requires every single person to be re-identified, then it can be inferred from the table that the performance of our system is 39.4%. In most cases, however, re-identification is used to measure the overall flow and the important issue is therefore to have an acceptable number of true positives and a low number of false positives, where especially the latter is clearly obtained in our system. For comparison a commercial re-identification system based on Wi-Fi signals from smartphones operates with a performance of approximately 50% [4].

Similar to others working on re-identification we also compute the CMC-curve to show the recognition performance for different rank values, see Figure I.6. Each of

	A	B	C	D	E
Run 1	35	10	16	0	9
Run 2	34	12	12	1	11
Run 3	33	13	13	1	10
Run 4	34	12	15	1	8
Run 5	34	10	13	2	11
Average	34	11.4	13.8	1	11
Percentage			93.2%	6.8%	

Table I.1: Re-identification results.

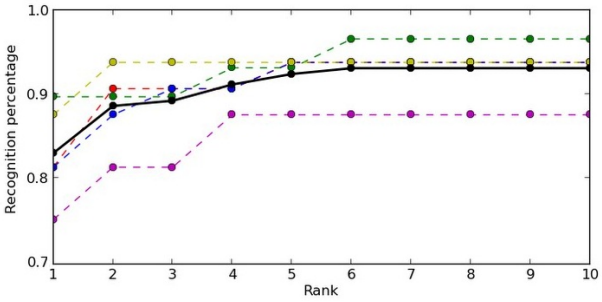


Fig. I.6: CMC-curve performance.

the dashed lines is a CMC-curve for a single run. The thick black line is the mean CMC of the 5 runs.

Since this is the first work on tri-modal re-identification we cannot compare our results directly with those of others. Instead in Table I.2 we list the rank-1 results of previous works. Please note that very different datasets and setting were used in these works and that no final conclusions therefore can be drawn. The results, however, seem to indicate the quality of our tri-modal approach, especially since we do not have a training phase as most others do.

7 Concluding remarks

We proposed a tri-modal re-identification system based on RGB, depth, and thermal descriptors. Three modalities were aligned, and robust discriminative features codifying soft biometrics were computed. The modalities were combined in a late fusion

Work	[1]	[2]	[6]	[7]	[8]	[10]	[14]	[21]	[23]	Our
Data	RGB	Depth	RGB	RGB	RGB	RGB	Thermal	RGB-D	RGB	RGB-D-T
Rank-1	51%	12%	N/A	82%	67%	43%	98%	78%	26%	82%

Table I.2: Data types and rank-1 results of recent re-identification works. Note that several works test on a number of different settings and different datasets. In such cases the table contains the average of the best results.

fashion, being able to predict a new user in the scene as well as to recognize previous users based on a combined rule cost. We tested our tri-modal re-identification system on anovel tri-modal dataset. Our results showed that the combination of all three modalities is the one that achieved better performance. A place to improve the system is in the determination of new persons. Nearly all new persons are detected as such, but there is a substantial amount of wrong New Persons. That is not a big issue with regards to re-identification performance, as presumably they will also be difficult to re-identify (they are only detected as new because they are not similar to the known persons), and in many applications it is not critical to be able to re-identify each and every subject. However, fewer wrong New Persons will result in a lower absolute re-identification rate.

Bibliography

[1] S. Bak, G. Charpiat, E. Corvée, F. Brémond, and M. Thonnat. "Learning to Match Appearances by Correlations in a Covariance Metric Space". In: *ECCV (3)*. Vol. 7574. LNCS. Springer, 2012, pp. 806–820. ISBN: 978-3-642-33711-6.

[2] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. "Re-identification with RGB-D Sensors." In: *ECCV Workshops (1)*. Vol. 7583. LNCS. Springer, 2012, pp. 433–442. ISBN: 978-3-642-33862-5.

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. "Speeded-Up Robust Features (SURF)". In: *Computer Vision and Image Understanding* 110.3 (2008). Similarity Matching in Computer Vision and Multimedia, pp. 346–359. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2007.09.014.

[4] Blip Systems. *Urban Planning*. http://www.bliptrack.com/urban/products/bliptracktm sensor/. 2013.

[5] G. Bradski and A. Kaehler. "Learning OpenCV". In: O'Reilly, 2008. Chap. 7, pp. 201–202.

[6] M. Demirkus, K. Garg, and S. Guler. "Automated person categorization for video surveillance using soft biometrics". In: *Biometric Technology for Human Identification VII*. 2010.

[7] Gianfranco Doretto, Thomas Sebastian, Peter H. Tu, and Jens Rittscher. "Appearance-based person reidentification in camera networks: problem overview and current approaches". In: *J. Ambient Intelligence and Humanized Computing* 2.2 (2011), pp. 127–151.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. "Person re-identification by symmetry-driven accumulation of local features". In: *CVPR*. 2010.

[9] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Vol. 2. Cambridge Univ Press, 2000.

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. "Person Re-identification by Descriptive and Discriminative Classification". In: *SCIA*. Vol. 6688. Lecture Notes in Computer Science. Springer, 2011, pp. 91–102. ISBN: 978-3-642-21226-0.

[11] B. Iglewicz and D.C. Hoaglin. *How to Detect and Handle Outliers*. ASQC basic references in quality control. ASQC Quality Press, 1993. ISBN: 9780873892476. URL: http://books.google.es/books?id=siInAQAAIAAJ.

[12] Kai Jüngling and Michael Arens. "A multi-staged system for efficient visual person reidentication". In: *Conference on Machine Vision Applications, Nara, Japan*. 2011.

[13] Kai Jüngling and Michael Arens. "Local Feature Based Person Reidentification in Infrared Image Sequences." In: *AVSS*. IEEE Computer Society, 2010, pp. 448–455. ISBN: 978-0-7695-4264-5.

[14] Kai Jüngling and Michael Arens. "View-invariant person re-identification with an Implicit Shape Model". In: *AVSS*. IEEE Computer Society, 2011, pp. 197–202. ISBN: 978-1-4577-0845-9.

[15] S.J. Krotosky and Mohan M. Trivedi. "On Color-, Infrared-, and Multimodal-Stereo Approaches to Pedestrian Detection". In: *Intelligent Transportation Systems, IEEE Transactions on* 8.4 (Dec. 2007), pp. 619–629. ISSN: 1524-9050. DOI: 10.1109/TITS.2007.908722.

[16] Chen Change Loy, Tao Xiang, and Shaogang Gong. "Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding". In: *Int. J. Comput. Vision* 90.1 (Oct. 2010), pp. 106–129. ISSN: 0920-5691. DOI: 10.1007/s11263-010-0347-5. URL: http://dx.doi.org/10.1007/s11263-010-0347-5.

[17] Brian E. Moore, Saad Ali, Ramin Mehran, and Mubarak Shah. "Visual crowd surveillance through a hydrodynamics lens". In: *Commun. ACM* 54.12 (2011), pp. 64–73.

[18] M. Robnik-Sikonja and I. Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning". In: *Machine Learning* 53 (2003), pp. 23–69.

[19] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. "Real-time human pose recognition in parts from single depth images". In: *CVPR*. IEEE, 2011, pp. 1297–1304.

[20] Berkan Solmaz, Brian E. Moore, and Mubarak Shah. "Identifying Behaviors in Crowd Scenes Using Stability Analysis for Dynamical Systems". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.10 (2012), pp. 2064–2070.

[21] C. Velardo and J. Dugelay. “Improving Identification by Pruning: A Case Study on Face Recognition and Body Soft Biometric”. In: *WIAMIS*. IEEE, 2012, pp. 1–4. ISBN: 978-1-4673-0791-8.

[22] Stephen Vidas, Ruan Lakemond, Simon Denman, Clinton Fookes, Sridha Sridharan, and Tim Wark. “A mask-based approach for the geometric calibration of thermal-infrared cameras”. In: *Instrumentation and Measurement, IEEE Transactions on* 61.6 (2012), pp. 1625–1635.

[23] W. Zheng, S. Gong, and T. Xiang. “Person re-identification by probabilistic relative distance comparison”. In: *CVPR*. IEEE, 2011, pp. 649–656.

ISSN (online): 2246-1248
ISBN (online): 978-87-7112-333-3

AALBORG UNIVERSITY PRESS